University of Siena – Department of Medical Biotechnologies

Doctorate in Genetics, Oncology and Clinical Medicine (GenOMeC)

XXXIII cycle (2017-2020)

Coordinator: Prof. Francesca Ariani

Sequencing-based approaches for the study of Lung-related diseases

Scientific disciplinary sector: MED/06 – Medical Genetics

Tutor

Dr. Silvestro Conticello

PhD Candidate

Dr. Filippo Martignano

FILIPPO
MARTIGNANO
04.05.2021
17:17:50 UTC

Academic Year 2019/2020

Preface

My thesis is focused on sequencing-based methods for lung diseases monitoring. Modern sequencing techniques allow us to comprehensively characterize nucleic acids obtained from patient-derived biological material, with potential applications in both basic research and clinical practice.

The thesis is divided in two sections:

In Section 1: “Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2” I describe the presence of RNA editing events in SARS-CoV-2, by analysing publicly available second generation RNA sequencing data from infected patients.

In Section 2: “Analysis of copy number variations from cell-free DNA of lung cancer patients via Nanopore sequencing” I have developed a customized workflow to exploit Nanopore sequence for the analysis of plasmatic cell-free DNA. The technique has been tested on plasma samples obtained from lung cancer patients, with the aim of detecting tumor-specific copy number variations. The approach has been subsequently validated by comparing it with the current standard technique (second generation sequencing: Illumina).

LIST of abbreviations

ACE-2: angiotensin-converting enzyme 2.

ADAR: Adenosine Deaminases acting on RNA.

APC: Antigen-presenting cell.

ApoB: Apolipoprotein B.

APOBEC1: Apolipoprotein B messenger RNA Editing Enzyme Catalytic Subunit 1.

ARDS: Acute Respiratory Distress Syndrome.

BALF: bronchoalveolar lavage fluid.

cfDNA: cell-free DNA

CGH: Comparative Genomic Hybridization.

CNV: copy number variation.

COVID-19: novel coronavirus disease 2019

ctDNA: circulating tumor DNA.

CTs: number of amplification cycles

ddPCR: Digital droplet PCR.

dNTP: deoxynucleoside triphosphate.

dsRNA: double stranded RNA.

E: envelope protein.

ERGIC: endoplasmic-reticulum-Golgi intermediate compartment.

HE: hemagglutinin esterase protein.

HIV: human immunodeficiency virus.

HLA: Human leukocyte antigen.

HMW: high molecular weight

LMW: low molecular weight.

M: membrane protein.

MERS-CoV: middle east respiratory syndrome coronavirus.

MHC: Major Histocompatibility Complex.

miRNA: micro-RNA.

MLPA: Multi Ligation-dependent Probe Amplification.

mRNA: messenger RNA.

N: nucleocapsid phosphoprotein.

NSP: nonstructural protein.

ONT: Oxford Nanopore Technologies.

ORF: open reading frame.

qPCR: Quantitative real-time PCR.

RBD: receptor binding domain.

RC: read count.

RdRp: RNA-dependent RNA polymerase

RTC: replicase-transcriptase complex.

S: spike protein.

SARS-CoV-2: severe acute respiratory syndrome coronavirus 2.

SARS-CoV: severe acute respiratory syndrome coronavirus.

SBS: sequencing by synthesis.

SGS: Second Generation Sequencing.

SNP: Single-Nucleotide Polymorphism.

SNV: Single nucleotide variation.

ssRNA: single-stranded RNA.

SV: structural variation.

SWGS: Shallow Whole Genome Sequencing.

WGS: Whole genome sequencing

Summary

Section 1: Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2

Introduction

Origin and characteristics of SARS-CoV-2

Emerging viral infections represent a threat to global health, and the recent outbreak of novel coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) exemplifies the risks (1, 2).

Coronaviruses are enveloped viruses, members of the subfamily *Coronavirinae* in the family *Coronaviridae* and the order *Nidovirales*; they have a positive nonsegmented single-stranded RNA (ssRNA) genome with a length ranging from 26 to 32 kb (3), which is structurally similar to eukaryotic messenger RNAs (mRNAs) in having 5'caps and 3′ poly-adenine tails (4). The coronavirus genome codes for membrane (M), spike (S), envelope (E) and hemagglutinin esterase (HE, not always present) structural proteins. These are responsible for cell infection/entry mechanisms and virion assembly, and are exposed on the surface of the virion giving it its distinctive "spiked" shape (Figure 1) (4). The structural nucleocapsid phosphoprotein (N) is on the other hand located inside the inner membrane of the virion, and it is associated with viral genomic RNA forming a

ribonucleoprotein with a helical structure (5, 6).

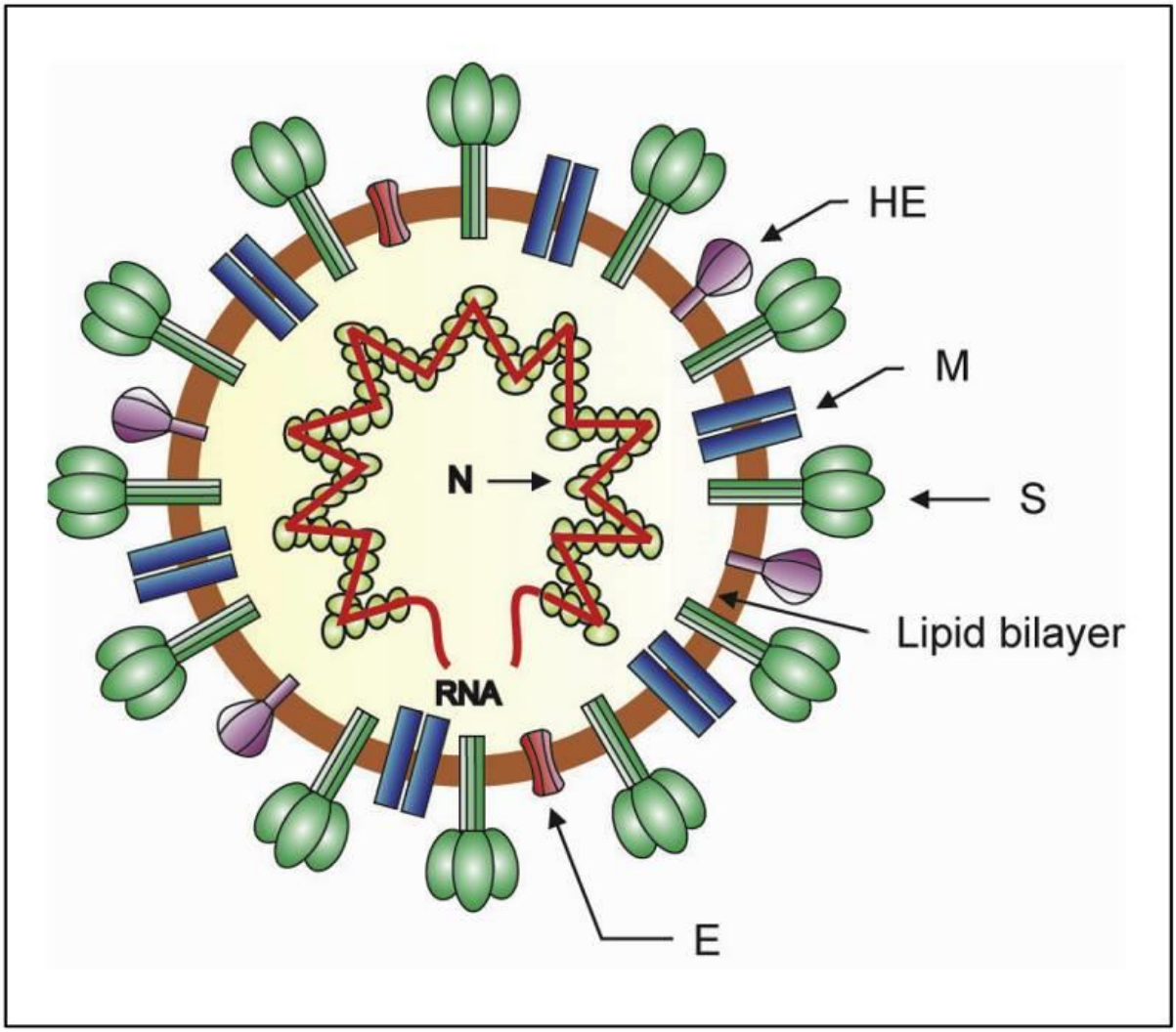

Figure 1: Stylized representation of coronavirus virion. Viral envelope is constituted by structural proteins M, S, H, E and HE, inserted in a lipid bilayer. Viral RNA genome is protected inside the envelope and associated with N structural proteins forming the helical ribonucleoprotein. Figure taken from (4) .

The 5'-most end of the genome is occupied by open reading frame (ORF) 1a and ORF1b, which constitute almost two-thirds of the entire region and code for 16 nonstructural proteins (NSPs) responsible of viral genes transcription and genome replication. Finally, coronaviruses possess a variety of accessory proteins whose number depends on the strain (Figure 2). Despite being dispensable, accessory proteins may confer biological advantages for the coronaviruses in the infected host cells (4).

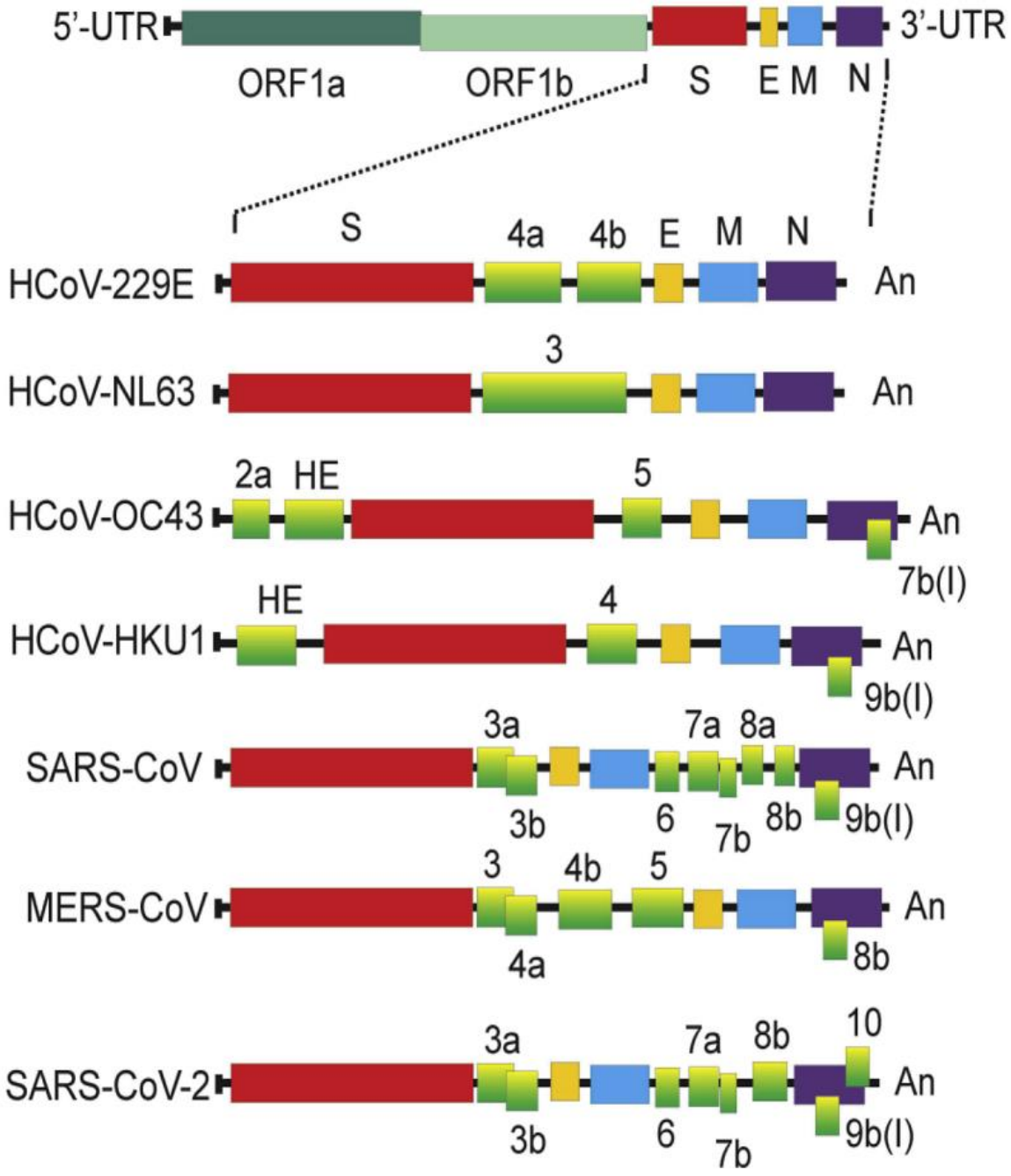

Figure 2: Schematic representation of human-infecting coronavirus genomes. Numbers and letters indicate structural, nonstructural and accessory proteins, Figure taken from (4) .

Despite being enveloped viruses, coronaviruses are far from being fragile or quickly inactivated; they are more robust than, for example, Human Immunodeficiency Virus (HIV)-1 and their infectivity can persist after 1-4 days on the relatively harsh environment of hard surfaces (4).

The genera *Alphacoronavirus* and *Betacoronavirus* are of particular interest due to their ability to infect mammals, usually causing respiratory illness in humans and gastroenteritis in animals (3). In particular, Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) are responsible respectively of the china outbreak in 2002–2003 and the Middle East outbreak in 2012 (7).

The emergence and re-emergence of coronaviruses is thought to be facilitated by increased contact of humans with wildlife (in particular in developing regions), accompanied often by lack of strict regulamentations, or local uses and costumes which encourage close contacts with natural reservoirs of novel viruses (4).

During the 2002 outbreak, the majority of early SARS-CoV cases were people attending chinese wildlife markets, in close contact with wild animals such as palm civets.

Subsequently, many coronaviruses phylogenetically related to SARS-CoV were discovered in bats from different chinese provinces; hence bats have been identified as the natural reservoir for SARS-CoV, with palm civet as the intermediate host. Spillover to humans was likely caused by multiple mutations acquired by the virus during infection in palm civets. Similarly, for MERS-CoV, bats are considered to be the natural reservoir and dromedary camels are the intermediate host (3, 8). Interestingly, most of the SARS-related viruses able to infect humans are found in China; it is therefore generally believed that bat-derived coronaviruses could re-emerge in future, causing outbreaks, with China being a likely hotspot (9).

SARS-CoV-2 is the last discovered member of the genus *Betacoronavirus* known to infect humans. Many theories have been proposed but, to date, the exact origin of SARS-CoV-2 is still a matter of debate (3, 7). Phylogenetic comparison of coronavirus sequences from the patients of different geographical regions, and climatic conditions supports the natural origin of SARS-CoV-2 (10-14).

Life-cycle and infection mechanisms of SARS-CoV-2

The first step of SARS-CoV-2 infection is the binding of viral extracellular S protein to Angiotensin-converting enzyme 2 (ACE-2) receptors on host cellular membrane, that provokes the fusion of viral and host cells' membranes (4, 15). Human coronaviruses often differ for Receptor Binding Domains (RBDs) of S protein, which binds to different human receptors; notably both SARS-CoV and SARS-CoV-2 bind to ACE-2, and their RBDs are nearly identical, furtherly supporting a close evolutionary relationship between the two viruses (16).

Another pivotal step for viral entrance, is the cleavage of S protein by Transmembrane Serine Protease 2: after the binding of ACE-2 with S protein, the latter exposes the fusion peptide that is close to the cleavage site, to finally achieve the fusion of the viral membrane with the cellular membrane (17, 18). Protein fragments resulting from the cleavage of S protein, are released in the extracellular space and serve as decoys for the inhibition of antibody-mediated neutralization, enhancing the chance of a successful infection (19).

After entry, the viral genome is released into the cellular cytoplasm and becomes available for transcription/translation, in a process termed "uncoating" (4, 15).

Subsequently, coronavirus takes advantage of host translational machinery to translate the polycistronic gene ORF1, directly from the positive sense viral genomic RNA.

The translation of ORF1, which is composed by ORF1a and ORF1b can generate two polyproteins: pp1a resulting from the canonical translation of ORF1a, and pp1ab resulting from a minus 1 (-1) ribosomal frameshift which bypasses ORF1a stop codon, leading to a fusion protein including both ORF1a and ORF1b (20).

The cleavage of pp1a/pp1ab generates 16 NSPs; among them, the proteases NSP3 and NSP5 are responsible for the autoprocessing of pp1a/pp1ab itself (21, 22).

Most of the NSPs assemble into the replicase-transcriptase complex (RTC) which creates a favorable environment for viral RNA synthesis. For this purpose, NSP3, NSP4 and NSP6 exploit endoplasmic reticulum (ER) membranes to produce vesicles, to which RTC is bound via their transmembrane domains (23). Viral genome and RNA synthesis factors (including RTC and hundreds of hijacked host proteins) are concentrated in these organelle-like vesicles, which protect the virus from host defence mechanisms and exonucleases (24, 25). NSP1 is not included in such vesicles due to its ability to hamper translation by interacting with 40S ribosomal subunit and causing premature mRNA degradation. Indeed, coronavirus exploit NSP1 to hamper the translation of host mRNAs while redirecting the translation machinery towards the production of viral proteins (26).

NSP12 contains the RdRp domain, responsible for viral RNA transcription producing both genomic and (smaller) subgenomic RNAs. The transcription involves the production negative strand intermediates which are used as a template for the transcription of positive strand RNAs, and represent only the 1% of the total viral RNA. Genomic RNAs are exact copies of the viral genome which will constitute the viral progeny; while subgenomic RNAs are portions of the viral genome, sharing the 3' end, used as mRNA for the translation of structural and accessory proteins. After the translation, structural proteins S, E and M are inserted into RE membrane and, subsequently, reach the endoplasmic-reticulum-Golgi intermediate compartment (ERGIC). Finally, viral genome is encapsidated by N protein and included into ERGIC membranes, forming mature virions which are transported to the cell surface and released by exocytosis (4, 27).

Evolution of SARS-CoV-2

The accumulation of mutations and homologous/nonhomologous recombination events (occurring in intermediate hosts and natural resevoirs) are decisive factors linked to the ability

of viruses to cross the species barrier and, in this context, to affect humans (27-30). The nature of viral genetic material is an important factor with regard to propensity for emergence. Roughly 85% of emerging viruses posses ssRNA genomes, and this may be related to their mechanism of replication which is highly error-prone. Indeed, the error rate of RNA genome replication is generally about 10^{-4}, and order of magnitude higher than DNA viruses ($\sim 10^{-5}$). Such high error rate is due to RNA polymerase, responsible of viral replication, which lacks the proofreading and post-replication mismatch repair features, fuelling ssRNA viruses' predisposition to mutate and evolve (4, 31, 32). Due to its strand switching ability, viral RNA polymerase is responsible also of homologous and nonhomologous recombination events (27). Notably, in Coronaviruses, Nsp14 mediates a form of error correction which helps reducing the overall mutational rate (33).

SARS-CoV-2 shares ~75–80% of its viral genome with SARS-CoV; and its genome 96% identical to the bat SARS-like coronavirus strain BatCov RaTG13 genome. This suggests that, once again bats, are likely to be reservoir hosts for this strain (8). Currently, the most likely intermediate host is the Malytan pangolin: an illegally trafficked species which is very popular in China for traditional medicine. Pangolin-derived coronavirus samples show a 85.5-92,4% homology with SARS-CoV-2 (34-36).

With regards to single viral proteins: SARS-CoV-2 and SARS-CoV S proteins show ~77% identity in the aminoacidic sequence (37-39).

Furthermore, the S protein RBD of SARS-CoV-2 and pangolin coronavirus are extremely close in terms of sequence similarity (99%) (36, 40). Such evidences suggest that SARS-CoV-2 may be the result of the recombination of two viruses, without any trace of human-mediated genetic manipulation.

SARS-CoV-2 genomes isolated from different patients show more than 99.9% sequence identity, suggesting a very recent host shift of this virus to humans (12, 14, 41).

According to a phylogenetic network analysis of 160 complete human SARS-CoV-2 genomes, it is already possible to define three main variants (A, B and C, with A being the ancestral type according to the bat outgroup coronavirus. Such variants have been defined) on the basis of aminoacidic changes. The A and C types belonged to the Europeans and Americans while the B type is the most common type in East Asia (42).

Pathogenesis and host response

The pathogenesis of SARS-CoV-2 is currently under the spotlight of a large section of the scientific community. However, SARS-CoV-2 specific studies are needed since most of our knowledge still derives from previous studies regarding similar viruese such as SARS-CoV and MERS-CoV. SARS-CoV-2 typically infects epithelial cells of the upper respiratory tract (i.e. oral an nasal cavities) which represent the first site of viral replication; during the disease progression it eventually reaches the conducting airways, where it infects primary ciliated cells. Most of the patients (~80%) have a mild course limited to upper and conducting airways. Alternatively, similarly to SARS-CoV, the virus can proceed infecting alveolar type II pneumocyte cells which comprise 10-15% of total lung cells and are responsible for the maintenance of surface tension in alveolar walls by producing surfactant.

Those cells are also important players in the maintenance of the lung epithelium after injury through epithelial regeneration. SARS-CoV-2 infection cause apoptosis of alveolar type II pneumocyte leading to serious injury of the lungs, impairing gas exchange which is hypothesized to lead to Acute Respiratory Distress Syndrome (ARDS) (7, 43, 44).

Intestinal enterocytes are another possible target of SARS-CoV-2 infection, which, in a subset of patients, can cause gastrointestinal symptoms (45). Notably, ACE-2 receptor is highly expressed in both enterocytes and pneumocytes, making these cellular types the preferred targets of infection (45). ACE-2 is a strong discriminant to determine the infectability of

human cells; indeed, Jia et al. reported the ability of SARS-CoV-2 to infect also adipose cells (which express ACE-2) (46). Also, ACE-2 expression is often reduced infected lung cells; its downregulation is associated with acute lung injury probably contributing to the development of ARDS (45, 47, 48).

During the infection, the activation of the body's humoral and cellular immunities is mediated by virus-specific B and T cells. In particular, studies on SARS-CoV show that cytotoxic T lymphocytes recognize viral antigens presented mainly via class I Major Histocompatibility Complexes (MHC) on Antigen-Presenting Cells (APCs). Different Human Leukocyte Antigen (HLA) genotypes are possibly linked to differences in susceptibility to the virus. HLA-B*46:01 allele has been associated with more severe manifestations of SARS-CoV infection; however, this relationship it has not been assessed yet with regard to SARS-CoV-2 (7, 49, 50).

On the other hand, innate immune system against coronaviruses is activated thanks to the recognition of viral genome fragments by toll like receptors 3 and 7, cytosolic RNA sensor, and RIG1/MDA5. Dendritic cells are widespread in the respiratory mucosa and are among the main contributors to innate response by producing type I IFNs and IL-6; also, they can serve as APCs to trigger adaptive immunity (45).

Immune system activation is characterized by a massive production of pro-inflammatory cytokines such as TGFβ, TNF-α, IFN-γ, IFN-α, IL-1β, IL-6, IL-8, IL-12, IL-18, and IL-33. The aim of cytokine production is the restrain of viral infection; however, the excessive and uncontrolled production of cytokines, termed “cytokine storm”, has deleterious effect on the patient (45). Indeed, the high levels of type I IFN, IL-2, IL-6, IL-7, IL-8, IL-10, MIP-1A, IP-10, G-CSF, MCP-1, and TNF-α has been associated to the to the progression of mild inflammation to severe inflammation in critical patients (51-53). The cytokine storm causes lung-tissue damage by activating the immune inflammatory cells to attack the alveoli and

produce fibrotic tissue in the lung. Also, it can lead to multiple-organ failure, which aggravates the health status of the patients, involving dysfunction of the kidneys, liver, heart, and other end organs (54).

RNA editing

RNA editing is a cellular mechanism involving post-transcriptional RNA modifications that cause single nucleotide variations (SNVs) in the mature transcript.

With regards to mRNAs, RNA editing can have a recoding function creating novel start/stop codons (55-57) or open reading frames (58); while editing of transfer RNAs can affect their function and structure (59-61).

RNA editing typically involves endogenous RNAs but, if targeting viral RNA, it is potentially deleterious for virus' viability itself, by generating premature stop codons and missense mutations in the viral genome. On the other hand, RNA editing on positive strand genomic RNAs, could fuel virus evolution by increasing the basal mutational rate. With regards to negative strand intermediate RNA: it is possible that the presence of edited bases leads to base mis-incorporations by the RdRp, which result in mutations in the progenie. However, to my knowledge, there are still no evidences that edited bases affect base incorporation specificity of RdRps (as they do with canonical DNA polymerases).

Two deaminase enzymes are responsible of RNA editing in higher eukaryotes:

- Apolipoprotein B messenger RNA Editing Enzyme Catalytic Subunit 1 (APOBEC1) (62): APOBEC1 catalyses the deamination of Cytidine to Uridine (C-to-U) on single stranded RNA (ssRNA), with cytosine 6666 of Apolipoprotein B (ApoB) mRNA as the main canonical target. Editing on ApoB happens only in the small intestine, causing the formation of a premature stop codon and leading to the correct maturation

of ApoB's mRNA (55, 56). For years, ApoB has been considered the only target of APOBEC1; more recently, additional targets have been identified: Neurofibromatosis type 1 in human peripheral nerve-sheath tumors (63), N-Acetyl-Transferase 1 in mouse and rabbit livers (64), and hundreds of transcripts in murine immune cells where APOBEC1 is strongly expressed (65-68).

Besides recoding functions, the effect of RNA editing on transcripts is not completely understood; it has been reported that it can affect mRNA fate by modifying micro-RNAs (miRNAs) binding sites (68).

APOBEC1 belongs to a larger family of deaminases (comprising AID, APOBEC1, APOBEC2, and APOBEC3 subgroups) (69).

APOBEC3A and APOBEC3G are the only other members of the APOBECs family able to edit RNA; most of their targets have been identified in white blood cells, and are involved in viral restriction pathways (70-72). Notably, the APOBEC3 sub-family is closely related to viral restriction; they have been proved effective against many viral species in experimental conditions, yet, until now, their mutational activity in clinical settings has been shown only in a handful of viral infections (73-80) through DNA editing. To date, the only evidence of RNA editing in viruses regards rubella virus (81).

- Adenosine Deaminases acting on RNA (ADAR) (82): The ADAR proteins catalyse the deamination of Adenine to Inosine (A-to-I) in double stranded RNAs (dsRNAs) via a hydrolytic mechanism (82-85). The catalytically active proteins ADAR1 and ADAR2 are ubiquitously expressed in all vertebrate tissues (86), with higher levels in the brain (87), where ADAR-mediated editing regulates neural signaling via recoding

of neurotransmitter receptors' and ion channels' mRNA (88). Only the 1% of human A-to-I editing sites affect coding regions (89); indeed, most of the targets include:

A) Non coding repetitive elements (90-92), with a possible role in transposable elements restriction.
B) miRNAs: affecting their specificity and the efficiency of their processing (93-96).
C) Introns and untranslated regions: affecting transcript stability (97-101).

ADARs' relationship with viral infections is quite contradictory: on one hand ADAR-related editing has been found in RNA viruses such HIV, Epstein-Barr and herpes virus (102, 103), potentially hampering viral integrity; on the other hand, they seem to have an inhibitory effect on immune system activation against exogenous dsRNA (104-106).

Considering the relationship between deaminases and immunity, and their ability to target viral genomes, it would not be surprising if they also play a role in coronavirus restriction via RNA editing. Indeed, expression of APOBEC3s is induced by mediators of inflammation, possibly reflecting their role as a first line of defense against invading viruses. In particular type I IFNs have been reported to enhance the expression of APOBEC3A, and APOBEC3G in monocyte, macrophages and plasmacytoid dendritic cells. Such IFN-mediated induction is mainly related to TLR activation. The massive production of cytokines, including type I IFNs, during coronaviruses infection may lead to APOBECs activation, along with their RNA editing activity. Also ADAR1 expression can be triggered by type I; in particular possesses an IFN-inducible variant (p150), which is induced following detection of viral infection (107). It has been reported that A-to-I RNA editing can facilitate TLR7/8

sensing of phagocytosed viral RNA (108). On the other hand, ADAR1 plays a role in avoiding overproduction of IFN by competing with RIG1 for the binding of exogenous RNA, and consequently preventing its activation (109, 110). These evidences suggest an active role of host deaminases during coronavirus infection, but it still has to be demonstrated if, in this context, their activation is associated with RNA editing activity on SARS-CoV-2

Sequencing-based detection of RNA editing

Second Generation Sequencing (SGS) is the technology of choice for the study of RNA editing, with Illumina being the leading company in the field.

Thanks to its high throughput and low error rate (~0.24%) (111), SGS allows de-novo detection of rare mutational events without prior knowledge of their genomic position, which make it particularly indicated for the study of noncanonical editing sites and off targets.

Illumina sequencers are based on the sequencing by synthesis (SBS) technology: a library of DNA fragments (genomic DNA or cDNA) is bound to a physical support (flow-cell); a sequencing cycle is composed by 4 steps, in each step a different fluorescently labeled deoxynucleoside triphosphate (dNTP) is added to the flow cell and incorporated in the growing filament. These dNTP are reversible terminators hence, for each template, only a single dNTP is added at the end of a sequencing cycle, and the base-specific fluorescence is detected photographically.

Before the start of a new sequencing cycle, terminator dNTPs are cleaved to allow incorporation of the next base (111).

It is possible to sequence one or both ends of a single DNA fragment (termed, respectively, single-end and paired-end sequencing) and a typical Illumina run is composed of 150 cycles for each fragment end (resulting in 150 bp long reads).

As a result, each read (or each couple of paired-end reads) represents part of the sequence of an input DNA fragment.

In a process called alignment (or mapping), the reads are compared with a reference sequence to identify the reference position they belong to, based on the degree of similarity between a candidate portion and the read itself: the higher the similarity (the lower the mismatches) the higher the probability the read is assigned to the correct position (112).

Once determined the most likely reference position, any eventual mismatch between the reference and the read represents a SNV (a mutation or, in this particular context, a RNA editing event). This concept is at the base of the so called "callers": tools for the detection of SNVs from Illumina reads, which mostly differ in filtering approaches for the removal of false positive calls (typically sequencing errors). The filtering strategy is usually based on the application they have been designed for: germline single nucleotide polymorphisms (SNPs), somatic mutations, editing events etc… (113-117)

The reliability of SNV calling strictly depends on sequencing coverage: coverage is the count of reads that include a specific reference position; in other words, it's the number of times a position has been sequenced and, hence, observed. It is then intuitive that, the higher the coverage the higher the accuracy.

The concept of coverage is closely related to the concept of "allelic fraction" (AF) which is calculated as following:

$$AF = ALT/COV$$

Where ALT is the number of reads carrying the mismatch (i.e. the alternative allele) and COV is the coverage in that specific position.

As sequenced reads are a proxy of input DNA fragments, the percentage of reads carrying the alternative allele represents the abundance of DNA carrying the SNV, allowing a quantitative

analysis (115, 118). With regards to RNA editing, AF calculation is of great importance as it can be interpreted as the frequency of an editing event and, as a rule of thumb, allows to discriminate an acquired SNV (typically low AF) from a germline SNV (typically AF ~ 50-100%).

Rationale

The aim of the study is to investigate ADAR and APOBEC-induced RNA editing on the coronavirus genome during infection in humans. From public repositories, we downloaded transcriptomic Illumina data obtained via RNA-sequencing of bronchoalveolar lavage fluid (BALF) samples from coronavirus infected patients. BALF is a diagnostic method of the lower respiratory system in which a bronchoscope is inserted in the lungs, with a measured amount of fluid introduced and then collected for examination (119). The fluid recovered is used to perform transcriptome analysis and has higher sensitivity compared to oropharyngeal and nasopharyngeal swabs (usually used for diagnostic purposes), for the detection of SARS-CoV-2 RNA (7, 120). It is hence the ideal technique to investigate viral genome sequences. We indeed detected putative RNA editing events from Illumina reads using two different softwares. Subsequently, we employed public genomic sequences of SARS, MERS and SARS-CoV-2 to assess the frequency of RNA editing events in the coronaviruses populations and the effect of both transcriptomic and genomic RNA editing events on protein traduction has been investigated.

A paper, including the following content, entitled “Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2”, DOI: 10.1126/sciadv.abb5813, is available at https://advances.sciencemag.org/.

Methods

Sequencing data

RNA sequencing data available from projects PRJNA601736, PRJNA603194, and PRJNA605907 were downloaded from the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/sra/) using the FASTQ-dump utilities from the SRA-toolkit with the following command line:

```
prefetch -v SRR* && fastq-dump --outdir /path_dir/ | --split-files
/path_dir/SRR*.sra
```

Table 1: Samples characteristics. * These samples were not considered because either the sequening depth was too low or the error rate was too high.

Run	BioProject	Library Selection	Instrument	Total reads	Mapped reads (%)	Mean coverage	Median coverage	Error rate	SNVs count
SRR10903401	PRJNA601736	RANDOM	Illumina MiSeq	953264	3.09	136.65	120	0.17%	25
SRR10903402	PRJNA601736	RANDOM	Illumina MiSeq	1353388	8.32	535.28	455	0.16%	163
*SRR10971381	PRJNA603194	RANDOM	Illumina MiniSeq	56565928	0.36	602.64	412	0.78%	NA
SRR11059940	PRJNA605907	RT-PCR	Illumina HiSeq 2500	79687	99.42	245.75	177	0.22%	24
*SRR11059941	PRJNA605907	RT-PCR	Illumina HiSeq 2500	13710	92.85	22.53	15	0.31%	NA
SRR11059942	PRJNA605907	RT-PCR	Illumina HiSeq 2500	2043855	99.85	6,991.56	2384	0.34%	208
*SRR11059943	PRJNA605907	RT-PCR	Illumina HiSeq 2500	190094	98.60	1,114.19	192	0.49%	NA
SRR11059944	PRJNA605907	RT-PCR	Illumina HiSeq 2500	1462225	99.11	4,345.56	2642	0.32%	111
SRR11059945	PRJNA605907	RT-PCR	Illumina HiSeq 2500	262312	98.21	578.24	53	0.41%	82
SRR11059946	PRJNA605907	RT-PCR	Illumina HiSeq 2500	7829225	99.59	22,582.02	12935	0.35%	238
SRR11059947	PRJNA605907	RT-PCR	Illumina HiSeq 2500	95405300	99.94	287,341.54	178543	0.29%	59

All the data has been produced by RNA-sequencing of BALF samples from SARS-CoV-2 infected patients. More details about the case series are available through the NCBI repository. Because most of the reads of samples from PRJNA605907 were missing their

mate, forward-reads and reverse-reads from these samples have been merged in a single FASTQ, which is treated as a single-end experiment. Details of the sequencing runs are summarized in Table 1.

Data preprocessing

SRR11059940, SRR11059941, SRR11059942, and SRR11059945 showed a reduced quality of the sequencing in the terminal part of the reads. We used TRIMMOMATIC (121) to trim the reads of those samples to 100 base bp, with the following command line:

```
rimmomatic SE SRR*.fastq SRR*.trimmed.fastq CROP:100
```

We aligned the FASTQ files using Burrows-Wheeler Aligner (112) using the official sequence of SARS-CoV-2 (NC_045512. 2) as reference genome. After the alignments, BAM files were sorted using SAMtools (113).

The command line used for paired-end samples is as follows:

```
bwa mem NC_045512.2.fa SRR*_1.fastq SRR*_2.fastq | samtools sort –O BAM -o SRR*_.bam
```

The command line used for single-end samples is as follows:

```
bwa mem NC_045512.2.fa SRR*.fastq | samtools sort –O BAM -o SRR*_.bam
```

The aligned bams have been analyzed with QUALIMAP (122). Because of a high error rate reported by QUALIMAP, samples SRR11059943 and SRR10971381 have been removed from the analysis.

SNV calling

A diagram of the entire pipeline is shown in Figure 3. We used REDItools 2 (116, 123) and JACUSA (117) to call the SNVs using the following command line:

```
python2.7 reditools.py -f SRR*.bam -o SRR10903401_stat_table_allPos.txt -S
-s 0 -os 4 -m /homol_site/SRR*_homopol.txt -c SRR*_homopol.txt -r
/Reference/NC_045512.2.fa -a SRR*_stat_table_allPos.txt -q 25 -bq 35 -mbp
15 -Mbp 15

jacusa call-1 -p 20 -r SRR*.vcf -a B,I,Y -s -f V -q 35 -m 25 SRR*.srt.bam
```

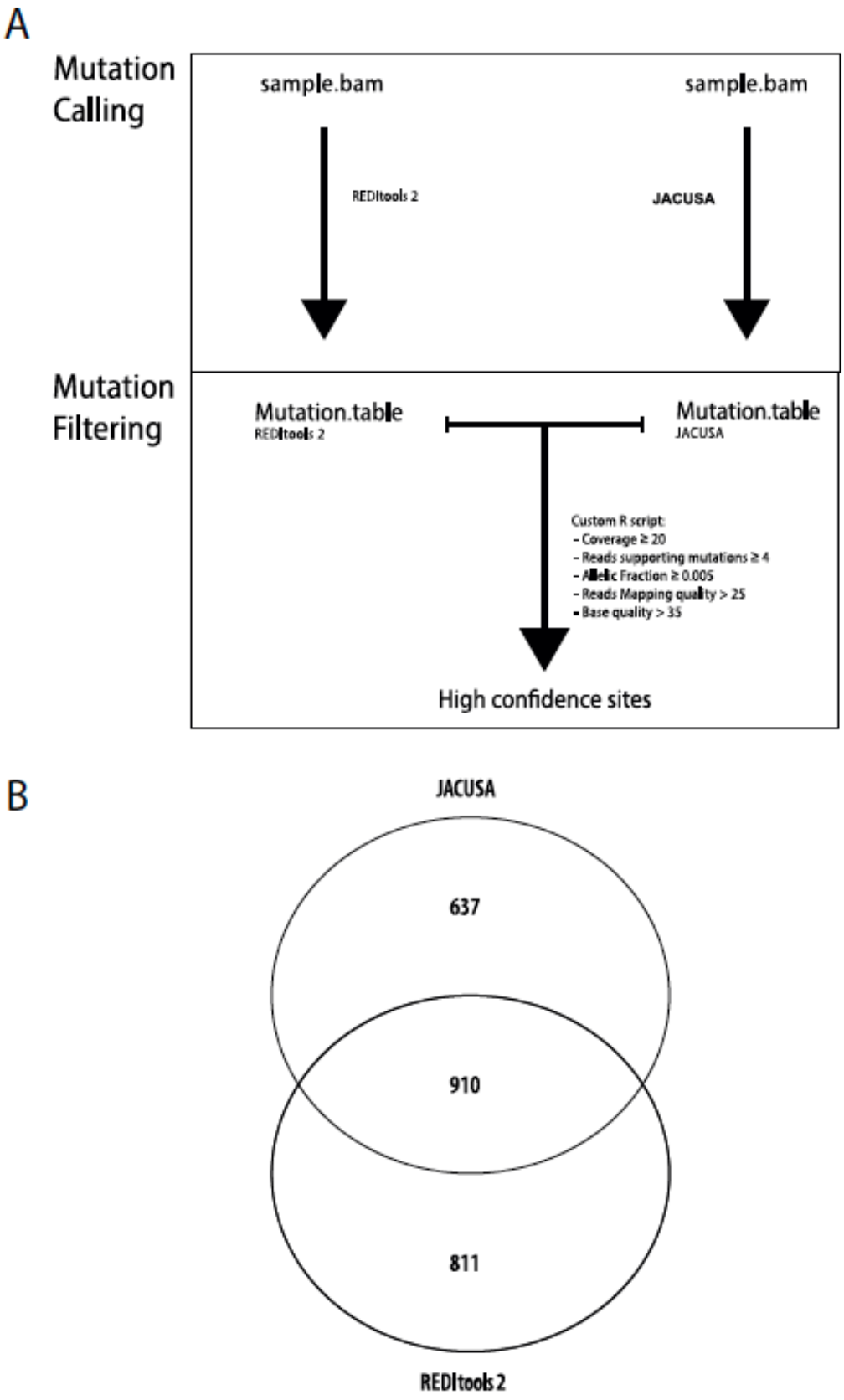

Figure 3: Schematic representation of the worflow for the detection of RNA editing events. (A) Mutation calling and filtering approaches via REDItools 2 and JACUSA. (B) Venn diagram of the SNVs identified by REDItools 2 and JACUSA.

With regard to REDItools 2, we removed all SNVs within 15 nucleotides from the beginning or the end of the reads to avoid artifacts due to misalignments.

To avoid potential artifacts due to strand bias, we used the AS_StrandOddsRatio parameter, calculated following GATK guidelines (https://gatk.broadinstitute.org/hc/en-us/articles/360040507111-AS-StrandOddsRatio), and any mutation with an AS_StrandOddsRatio > 4 has been removed from the dataset.

Bcftools (113) has been used to calculate total allelic depths on the forward and reverse strand (ADF and ADR) for AS_StrandOddsRatio calculation, with the following command line:

```
mpileup -a FORMAT/AD,FORMAT/ADF,FORMAT/ADR,FORMAT/DP,FORMAT/SP -O v -A -C -
I -d 1000000 -q 25 -Q 35 -f NC_045512.2.fa -o SRR*.vcf SRR*.srt.bam
```

Mutations common to the datasets generated by REDItools 2 and JACUSA were considered (n = 910; Figure 3). The percentage of concordant mutations doesn't depend on samples'

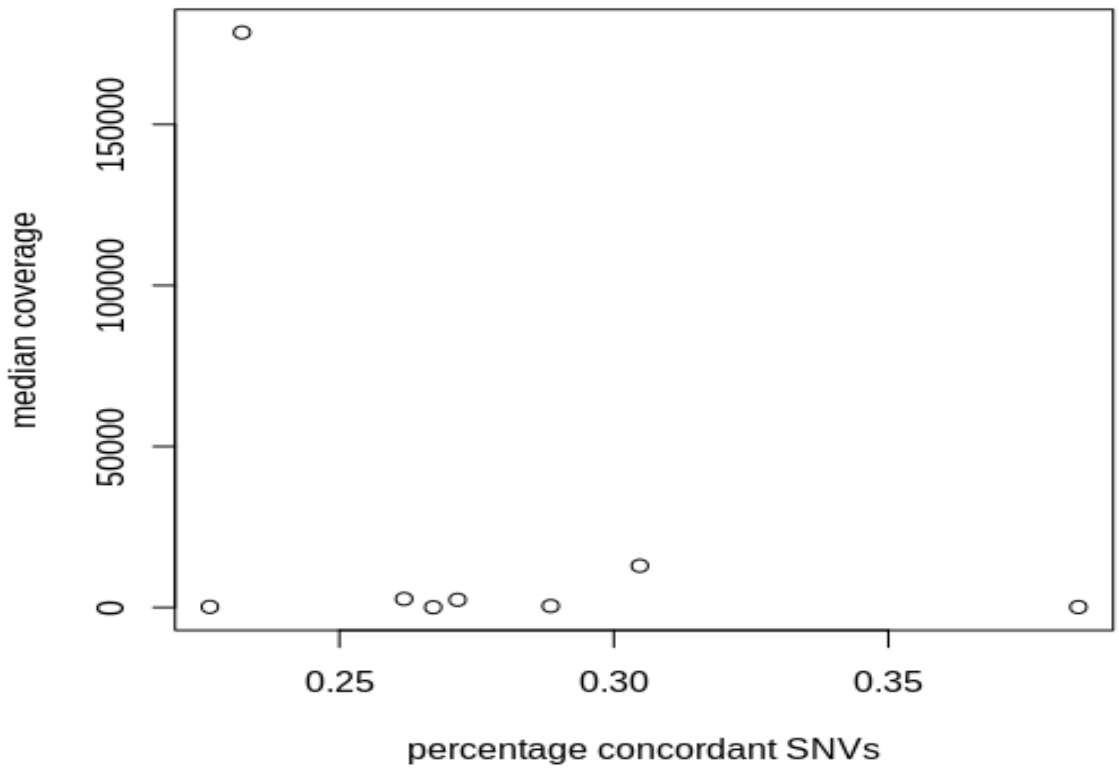

Figure 4: Relationship between samples' coverage and calling software concordance. Each sample is a dot, the percentage of concordance is calculated as follows: common SNVs/(Total Reditools2 SNVs + Total Jacusa SNVs).

coverage (Figure 4). The threshold we used to filter the SNVs is based on minimum coverage (20 reads), number of supporting reads (at least four mutated reads), allelic fraction (0.5%), quality of the mapped reads (>25), and base quality (>35). In the dataset, there were only six SNVs with allelic fractions in the range of 30 to 85% (C>T, 1; T>C, 3; G>T, 2). Because there were no SNVs with higher allelic fractions, we presume that all samples originated from the same viral strain. Recurring SNVs have been defined as the SNVs present in at least two samples. To overcome the problem of samples with lower sequencing depth, we used the positions of the SNVs common to both REDItools 2 and JACUSA to call again the SNVs irrespectively of the number of supporting reads.

Data manipulation

R packages (Biostrings, rsamtools, ggseqlogo ggplot2, and splitstackshape) and custom Perl scripts were used to handle the data.

Sequence context analysis

Logo alignments were calculated using ggseqlogo, using either the pooled dataset or the dataset of recurring SNVs. Logo alignments of the human edited sites were performed using ADAR sites from REDIportal (98) that were shared by at least four samples. SARS-CoV-2, SARS, and MERS genomic data were prepared for the Logi alignment using the GenomicRanges R package (124).

Normalized logo enrichment plots were generated with "two sample logos" (http://www.twosamplelogo.org/). Sequence contexts around Cs and As from reference genome were used as a control set when analysing respectively C-to-U and A-to-I editing sites.

SNV calling in genomic data from SARS-CoV-2, SARS, and MERS

The viral genomic sequences of MERS (taxid:1335626) and SARS (taxid:694009) were selected on NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) using the following query: Host : Homo Sapiens (human), taxid:9606; -Nucleotide Sequence Type: Complete. They were aligned using the "Align" utility. Consensus sequences of SARS and MERS genomes were built using the "cons" tool from the EMBOSS suite (http://bioinfo.nhri.org.tw/gui/) with default settings. SARS-CoV-2 genomic sequences were downloaded from GISAID (https://www.gisaid.org/) and aligned with MUSCLE (125).
SNVs have been called with a custom R script, by comparing viral genome sequences to the respective consensus sequence or, for SARS-CoV-2, to the NC_045512.2 reference sequence.

SNV annotation

SNVs (from both genomic and somatic SNV sets) occurring on coding sequences have been annotated with custom R scripts to determine the outcome of the nucleotide change (nonsense/missense/synonymous mutation). A summary is reported in Table 2.

Statistical analysis

fisher.test() function from the R base package has been used for all the statistical tests. To test the significance of C-to-U bias on the positive strand, we compared C>T/G>A SNV counts to the count of C/G bases on the reference genome. For *P* values of "RNA vs Reference," "DNA vs Reference," and "genome vs RNA," 2 × 2 contingency tables have been generated as shown in Table 2.

Results

To assess whether RNA editing could be involved in human host responses to SARS-CoV-2 infections, we started from publicly available RNA sequencing datasets from BALF obtained from patients diagnosed with COVID-19. While transcriptomic data for all samples could be aligned to the SARS-CoV-2 reference genome, the quality of the sequencing varied and only eight samples had coverage and error rates suitable for the identification of potentially edited sites (Table 1). We called SNVs on these eight samples (126, 127) using REDItools 2 (116, 123, 128) and JACUSA (117) using the following thresholds: reads supporting the SNV ≥4, allelic fraction ≥0.5%, coverage ≥20, quality of the reads >25, base quality >35 (Figure 3 **A**). The two pipelines gave comparable results with ~50% of the SNV positions called by both (Figure 3 **B,** Figure 5 and Figure 6). We identified 910 SNVs common to REDItools 2 and JACUSA, ranging from 24 to 238 SNVs per sample (Figure 7). Given the thresholds used to call the SNV, samples with lower sequencing depths displayed lower numbers of SNVs.

While the weight of each SNV type varies across samples (Figure 7), a bias toward transitions is always present, which is even more evident when all mutational data are pooled (Figure 8 **A** and **B**). This pattern holds true even when only SNVs recurring in more samples are considered (Figure 8 **C**).

The SNV allelic fraction (also referred to as frequency) and number of transversions are compatible with the mutation rates observed in coronaviruses [10–6/−7; (129)] and commonly associated to the RdRp. RdRps are error prone and are considered the main source of mutations in RNA viruses. However, the coronavirus NSP14-ExoN gene provides a form of error correction (33), which is probably the reason mutation rates in coronaviruses are lower than those observed in RNA viruses with smaller genomes. The mutational spectrum in SARS quasispecies presents a very weak bias toward U-to-G.

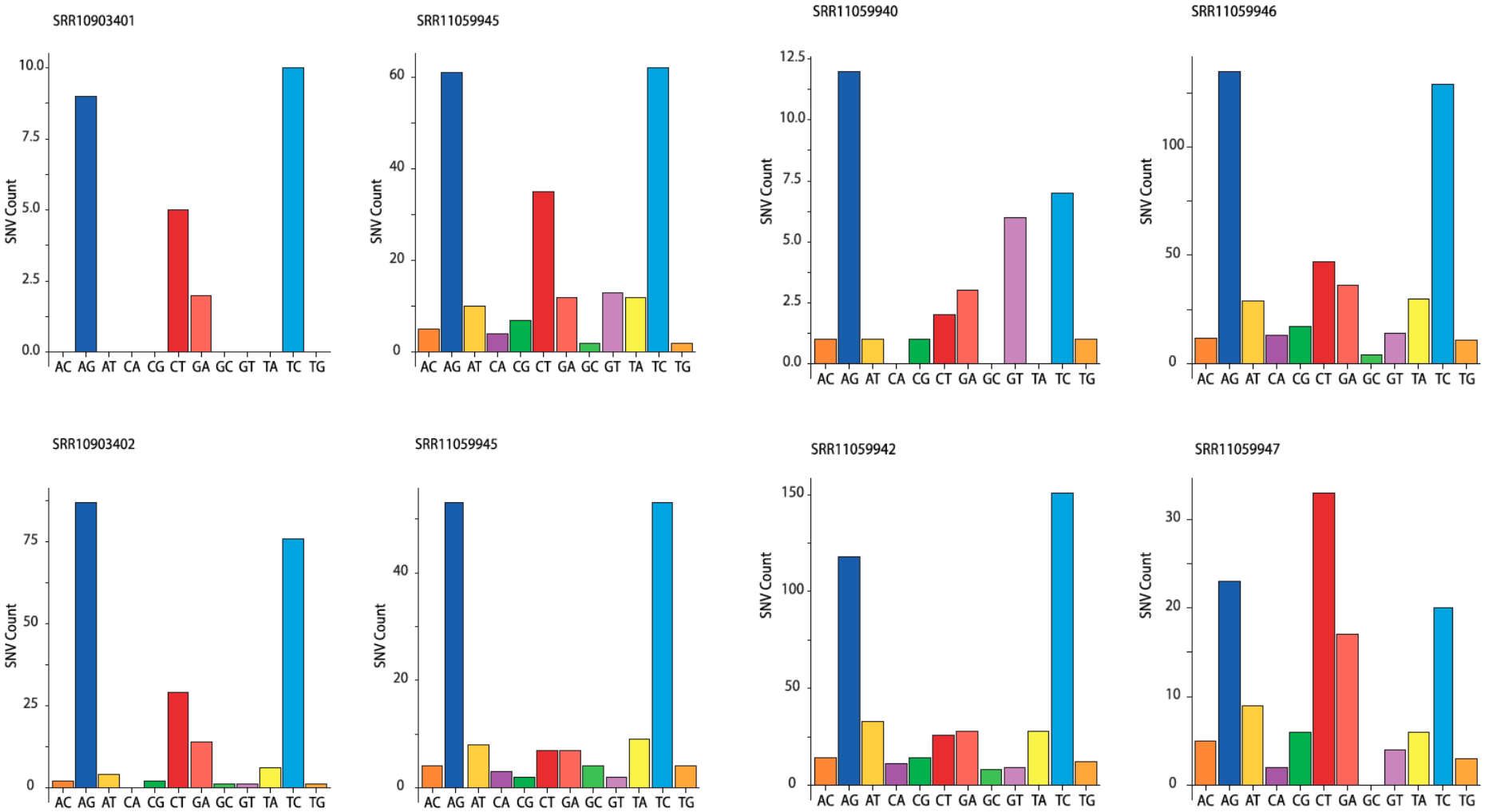

Figure 5: SNVs identified in SARS-CoV-2 transcriptomes by REDItools 2. The bar charts show the number of SNVs for each 2019-nCoV transcriptome (e.g. A>C, AC).

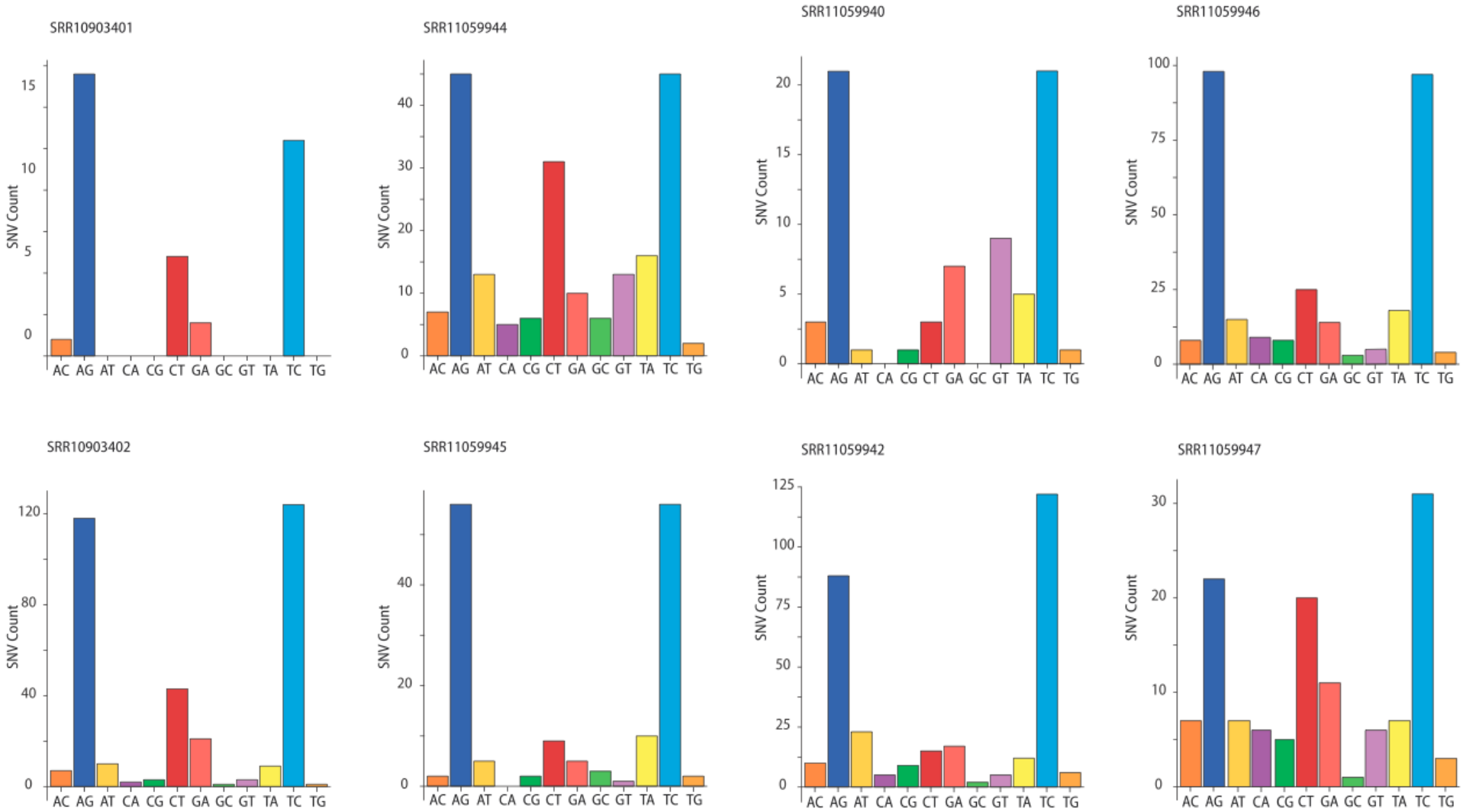

Figure 6: SNVs identified in SARS-CoV-2 transcriptomes by JACUSA. The bar charts show the number of SNVs for each 2019-nCoV transcriptome (e.g. A>C, AC).

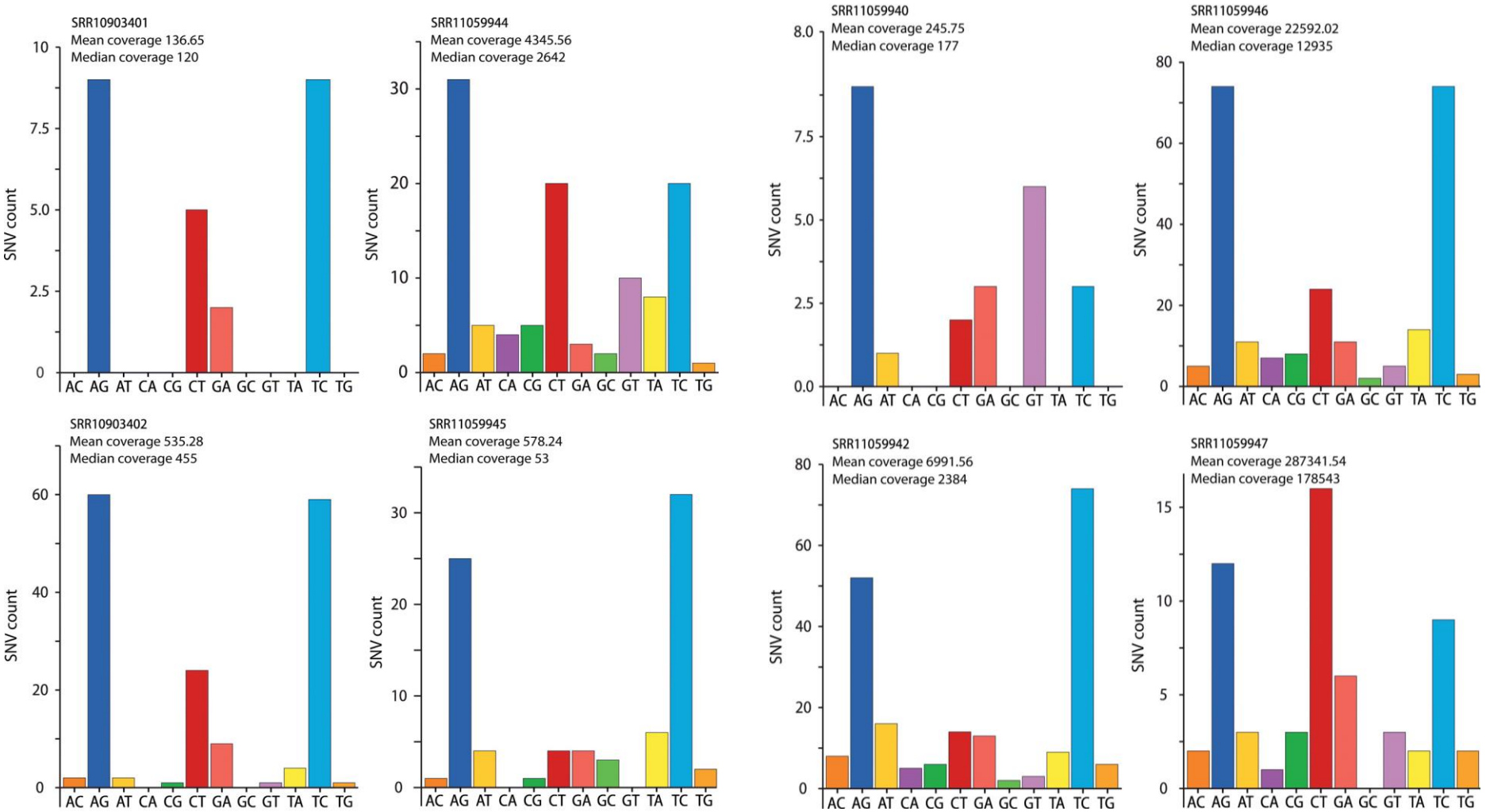

Figure 7: SNVs identified in SARS-CoV-2 transcriptomes after the intersection of Reditools 2 and JACUSA results. The bar charts show the number of SNVs identified in each SARS-CoV-2 transcriptome for each SNV type (e.g., A>C, AC). The sequencing depth for each sample is indicated.

Inactivation of NSP14-ExoN error correction reveals the mutational spectrum of the RdRp, which is quite different from the pattern we observe (i.e., main changes are C-to-A, followed by U-to-C, G-to-U, A-to-C, and U-to-G) (130). Hence, we would consider that SNVs deriving from RdRp errors represent a marginal fraction of the SNVs in the SARS-CoV-2 samples.

The bias toward transitions—mainly A>G/T>C changes—resembles the pattern of SNVs observed in human transcriptomes (97) or in viruses (131-133), where A>G changes derive from deamination of A-to-I mediated by the ADARs. It is thus likely that the A>G/T>C changes seen in SARS-CoV-2 are also due to the action of ADARs.

C>T and G>A SNVs are the second main group of changes and could derive from APOBEC-mediated C-to-U deamination. Unlike A-to-I editing, C-to-U editing is a relatively rare phenomenon in the human transcriptome (97), and with regard to viruses, it has been associated only with positive-sense ssRNA rubella virus (81), where C>T changes represent the predominant SNV type. The observation that only A-to-I editing is present in RNA viruses that infect nonvertebrate animals, where RNA-targeting APOBECs are not present (131, 132), supports the hypothesis that APOBECs are involved in the RNA editing of this human-targeting virus.

A third group of SNVs, A>T/T>A transversions, is also present in these samples. While this type of SNV has been reported in other genomic studies (134), its origin is still unknown.

A>G and T>C changes are evenly represented with respect to SNV frequency (Figure 8 **A**), the number of unique SNVs (Figure 8 **B** and **C**), and their distribution across the viral genome (Figure 8 **D**). As ADARs target dsRNA, this suggests that dsRNA encompasses the entire genome. While dsRNA in human transcripts is often driven by inverted repeats, the most likely source of dsRNA in the viral transcripts is replication, where both positive and negative strands are present and can result in wide regions of dsRNA.

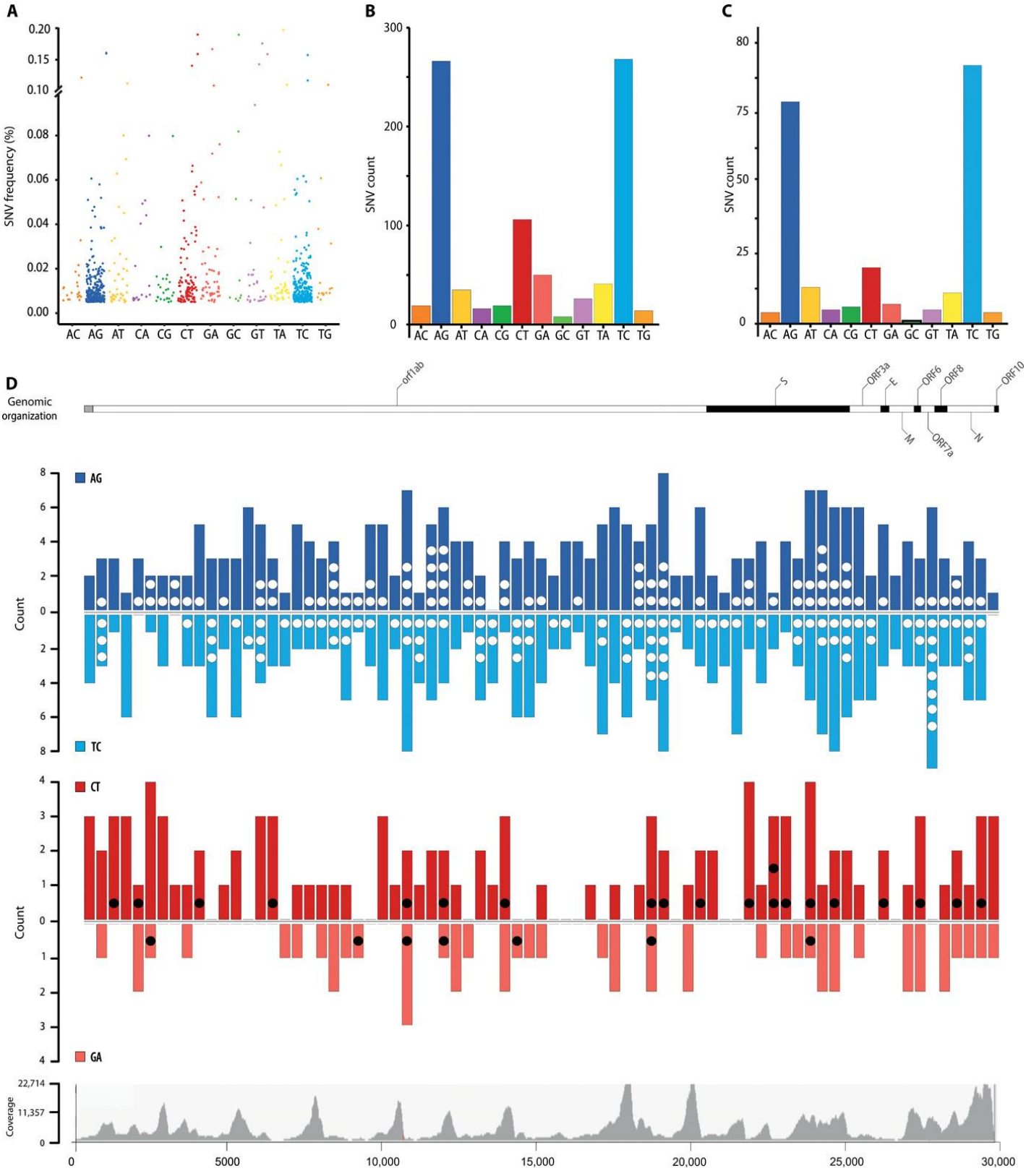

Figure 8: Total SNV identified in SARS-CoV-2 transcriptomes. (A) Allelic fraction and (B) number of SNVs for each nucleotide change in the entire dataset and (C) for SNVs recurring in at least two samples. (D) Distribution of SNVs across the SARS-CoV-2 genome. A-to-G (blue) and C-to-U (red) SNVs are grouped in 400-nucleotide (nt) bins and plotted above (AG and CT) or below the line (TC and GA) based on the edited strand. Dots (white/black) indicate recurring SNVs. Genetic organization of SARS-CoV-2 (top). The dark/white shading indicates the viral coding sequences; coverage distribution of all analyzed samples (bottom).

Unlike A-to-I changes, C-to-U changes are biased toward the positive-sense strand (Figure 8 **B to D**; $P < 0.0001$). Because ADARs and APOBECs selectively target dsRNA and ssRNA, this distribution could arise from the presence at all times of RNA in a dynamic equilibrium between double-strandedness—when negative-sense RNA is being transcribed—and single-strandedness—when nascent RNA is released. Although some areas seem to bear fewer SNVs, these reduced SNV frequencies might be related to lower sequencing depth in those regions.

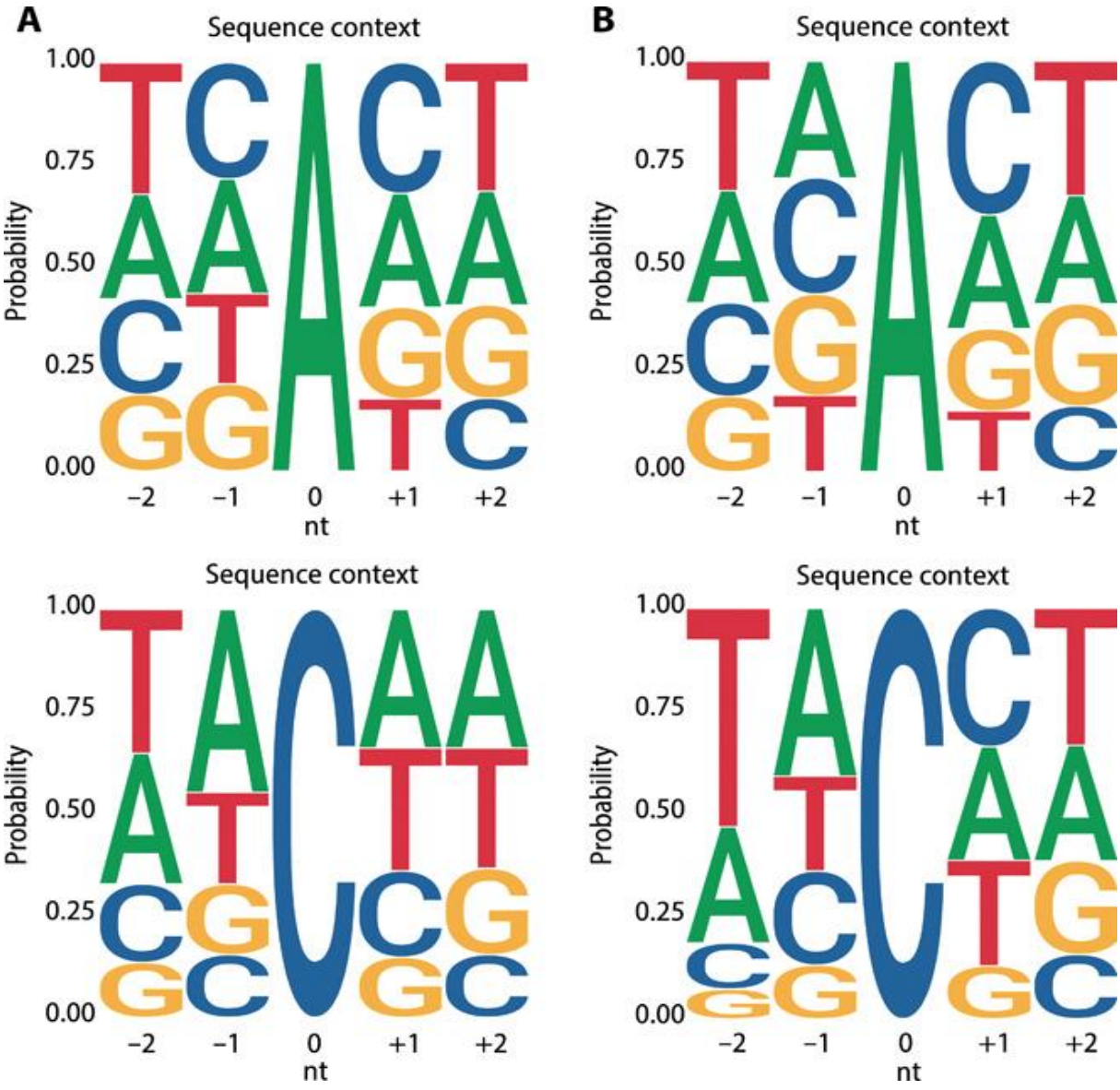

Figure 9: Logi alignment for SARS-CoV-2 RNA edited sites. (A) Local sequence context for A-to-I and C-to-U edited sites in the viral transcriptome and (B) for recurring sites.

As APOBEC deaminases preferentially target cytosines within specific sequence contexts, we analyzed the nucleotide context of A-to-I and C-to-U SNVs in the viral genome (Figure 9 **A** and **B**, Figure 10 **A**, **B**, **C**, **D**). A slight depletion of G bases in position −1 is present at A-to-I edited positions. This depletion is not as strong as the signal previously reported in human transcripts (91, 135-137). The low editing frequencies we observe resembles the editing present on human transcripts containing Alu sequences, which were found in a limited number in those early datasets. After the logi alignment, there is no evidence of a sequence context preference if we use a larger dataset such as REDIportal (98), which includes >1.5 M sites in Alu repeats (Figure 9). When normalising sequence contexts around A-to-I editing sites on sequence contexts around As in reference genome, a GC[A]S motif has been identified; however such motif is different than the ones reported in literature (91, 135, 137, 138).

With regards to the APOBECs, C-to-U changes preferentially occur downstream from uridines and adenosines, within a sequence context that resembles the one observed for APOBEC1-mediated deamination ([AU]C[AU]) (66, 139). However, no nucleotide enrichment was detected after sequence context normalization (Figure 10 **F**); raising the question of wether such sites could derive from random events rather than motif-specific mutations.

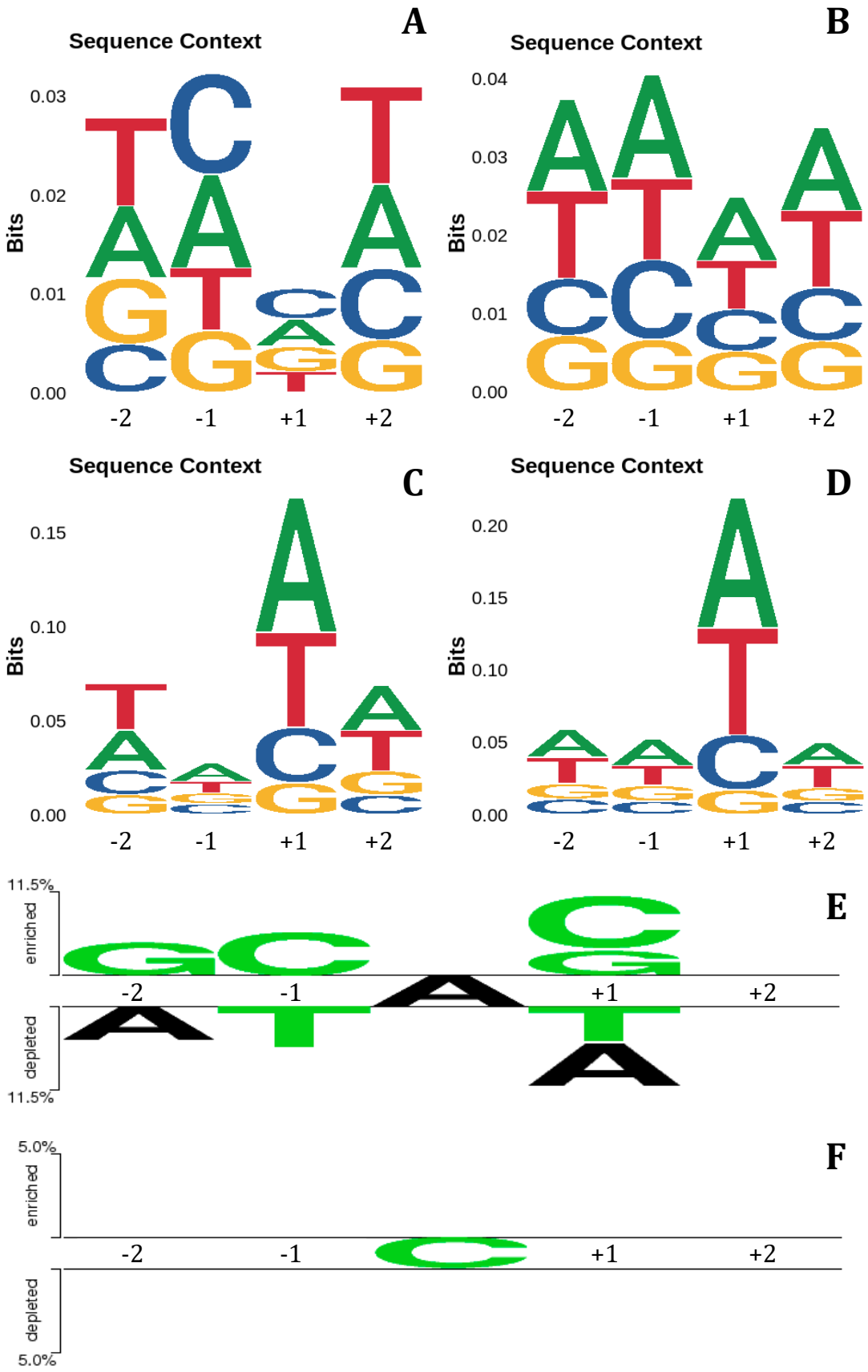

Figure 10. Additional logi alignments: logi aligments showing the information content of the position (bits) for (A) A-to-I edited sites, (B) reference genome sequence context around As, (C) C-to-U edited sites, (B) reference genome sequence context around Cs. Enrichment and depletion plot of sequence context of edited sites normalized against reference genome sequence contexts around candidate base for (E) A-to-I editing and (F) C-to-U editing.

We then aligned available genomes from SARS-CoV-2, Middle-East respiratory syndrome–related coronavirus (MERS-CoV), and SARS-CoV to test whether RNA editing could be responsible for some of the mutations acquired through evolution. The genomic alignments reveal that a substantial fraction of the mutations in all strains could derive from enzymatic deaminations (Figure 11 **A** to **C**), with a prevalence of C-to-U mutations, and a sequence context compatible with APOBEC-mediated editing also exists in the genomic C-to-U SNVs (Figure 11 **D** to **F**).

Discussion

Our data source—metagenomic sequencing—raises the question whether the low-level editing we observe (~1%) reflects the actual levels of editing of viral transcripts within human cells. Aside from a small fraction of cellular transcripts edited at high frequency, most ADAR-edited sites in the human transcriptome (typically inside Alu sequences) present editing levels of ~1% (97, 140, 141). It has been shown that a fraction of the cellular transcripts are hyperedited by ADARs (142-144).

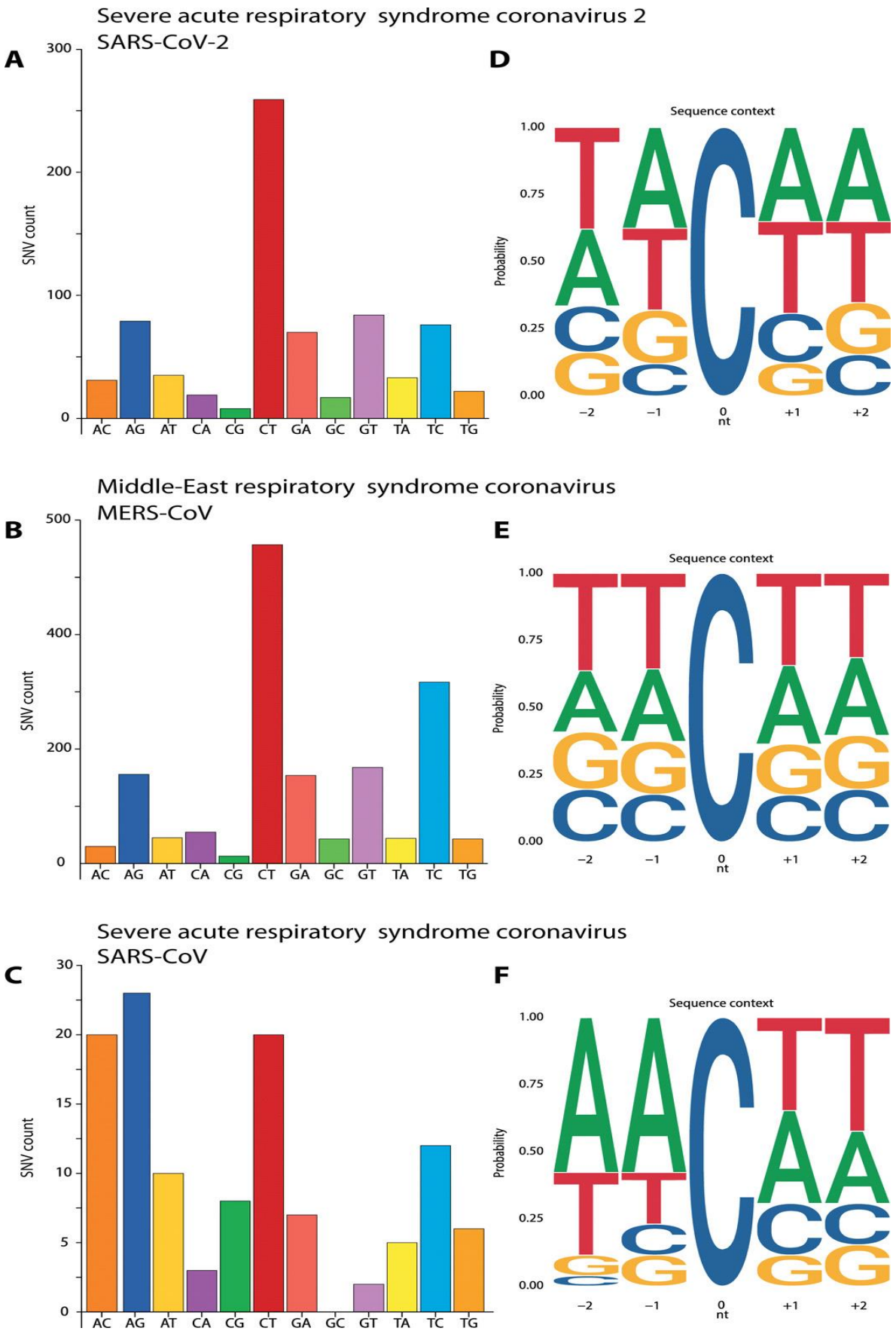

Figure 11: Nucleotide changes across Coronaviridae strains. (A to C) Number of SNVs for each nucleotide change and (D to F) local sequence context for C-to-U edited sites in genome alignments from SARS-CoV-2 (A and D), human-hosted MERS-CoV (B and E), and human-hosted SARS-CoV (C and F).

While we were unable to observe hyperedited reads in the metagenomic samples, it is possible that hyperedited transcripts fail to be packaged into the virus.

With regard to APOBEC-mediated RNA editing, its detection in the viral transcriptomes is already indicative, as this type of editing is almost undetectable in human tissues (97). This enrichment points either toward an induction of APOBECs triggered by coronavirus infection or to specific targeting of the APOBECs to the viral transcripts. APOBECs have been proved effective against many viral species in experimental conditions, yet, until now, their mutational activity in clinical settings has been shown only in a handful of viral infections (73-80) through DNA editing and, in rubella virus, on RNA (81).

Kim et al. (145) observed 41 recurrent base-modification sites in SARS-CoV-2 RNAs; it is possible that base modifications may lead to nucleotide misincorporation during PCR amplification, acting as a confounding factor and introducing biases in our analysis.

However, most of the base-modification sites fall approximately at 29000kb of viral genomic RNA (within N protein coding region), and we don't observe an enrichment of APOBEC and ADAR-related mutations in that region (Figure 8 **D**).

As in rubella virus, we observe a bias in APOBEC editing toward the positive-sense strand. This bias and the low editing frequencies might be indicative of the dynamics of the virus, from transcription to selection of viable genomes. It is reasonable to assume that sites edited on the negative-sense strand will result in a mid-level editing frequency, as not all negative-sense transcripts will be edited (Figure 12 **A**). On the other hand, editing of the positive-sense strand can occur upon entry of the viral genome, thus yielding high-frequency editing (Figure 12 **B**), or after viral genome replication, resulting in low-frequency editing (Figure 12 **C**). The lack of a sizable fraction of highly edited C>T SNVs suggests that APOBEC editing occurs late in the viral life cycle (Figure 12 **C**). Yet, because they occur earlier, G>A SNVs should be closer in number to C>T SNVs and with higher levels of editing, which is not what

we observe (Figure 8 **A** to **C**). The overrepresentation of C>T SNVs could be due to an imbalance toward positive-sense transcripts, as these are continuously generated from the negative-sense ones (and double-stranded hybrid RNAs are lost). However, the editing frequencies of G>A SNVs should be much higher, as G>A SNVs are generated upstream to the C>T ones. A more fitting explanation is that editing of the negative-sense transcripts results somehow in a loss of the edited transcript (Figure 12 **D**), lowering the chances of the edited site to be transmitted. Despite the fact, according to our data, that APOBEC-related mutations seem to be more deleterious than ADAR-related ones, it is possible that ADAR mediates hyper-editing (144), introducing a large amount of mutations on the same transcript. Such a load of mutations it is more likely to affect gene functionality rather than a single mutation, and it might be undetectable because of the loss of the edited (and not functional) transcript.

Taking in account our observation, with low editing frequency of both C>T and G>A mutations and with a higher number of unique C>T SNVs, the most likely compatible model is the one explained in Figure 12 **D**.

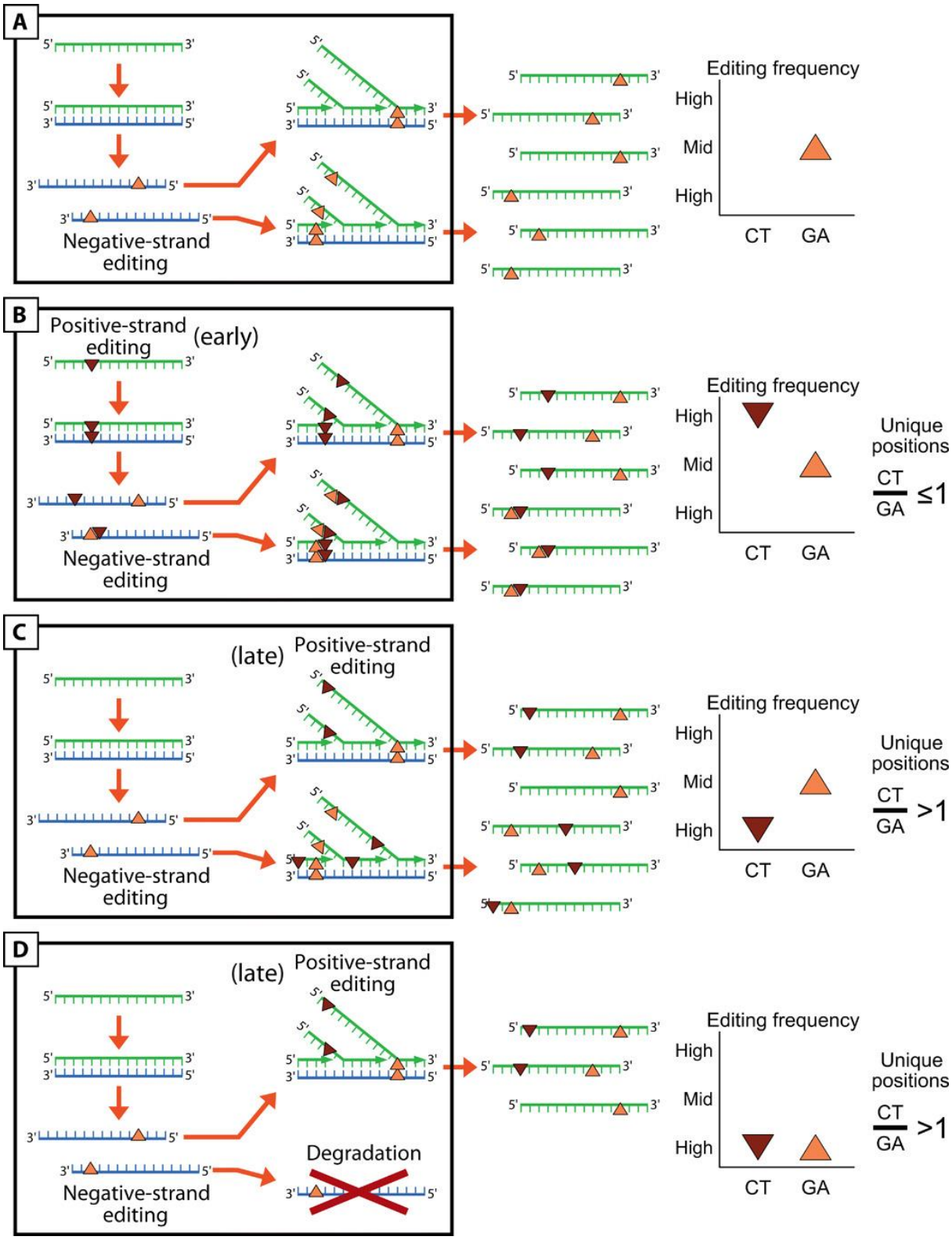

Figure 12: Model of APOBEC RNA editing on SARS-CoV-2 transcriptome. The four panels model the editing frequencies and the C>U/G/A ratios expected from four different scenarios: (A) C-to-U editing on the negative-sense transcripts, (B) "early" editing on the viral genomes before viral replication, (C) "late" editing after viral replication, and (D) "late" editing after viral replication with loss of negative-sense transcripts. Red dots indicate editing on the positive-sense transcript; orange dots indicate editing on the positive-sense transcript. Green and blue segments indicate positive- and negative-sense viral transcripts, respectively.

Because most of the APOBECs are unable to target RNA, the only well-characterized cytidine-targeting deaminases are APOBEC1, mainly expressed in the gastrointestinal tract, and APOBEC3A (70), whose physiological role is not clear. As with A-to-I editing, it will be important to assess the true extent of APOBEC RNA editing in infected cells.

The functional meaning of RNA editing in SARS-CoV-2 is yet to be understood: In other contexts, editing of the viral genome determines its demise or fuels its evolution. For DNA viruses, the selection is indirect, as genomes evolve to reduce potentially harmful editable sites [e.g., (131)], but for RNA viruses, this pressure is even stronger, as RNA editing directly affects the genetic information and efficiently edited sites disappear.

A comparison of the SNV datasets from the transcriptomic and genomic analyses reveals a different weight of A-to-I and C-to-U changes (Figure 8 **B** and Figure 11 **A**), with an underrepresentation of A-to-I in the viral genomes. As our analysis underestimates the amount of editing due to the strict parameters used, the underrepresentation of A-to-I changes could be explained by the possibility that A-to-I editing is more effective in restricting viral propagation, thus reducing the number of viral progeny showing evidence of these changes. In contrast, the remnants of less effective C-to-U editing are retained in viral progeny and get fixed during viral adaptation.

An analysis of mutation outcomes is difficult due to the low numbers of events collected so far, but there are some possibly suggestive trends (Table 2). C-to-U changes leading to stop codons are overrepresented in the transcriptomic data but—as expected—disappear in the genomic dataset. This might point—again—to an antiviral role for these editing enzymes. There is also an underrepresentation of C>T missense mutations, but its meaning is difficult to interpret.

	Reference: Potential SNVs on the virus				RNA: SNVs on the Transcriptomic dataset				genome: SNVs on the genomic dataset				P-values: RNA vs Reference			P-values: DNA vs Reference			P-values:
SNV type	Total	Stop	Miss	Syn	Total	Stop	Miss	Syn	Total	Stop	Miss	Syn	Stop	Miss	Syn	Stop	Miss	Syn	genome vs RNA
AC	6423	0	5269	1154	15	0	14	1	9	0	8	1	1,0000	0,4960	0,4960	1,0000	1,0000	1,0000	1,0000
AG	**6423**	**0**	**4524**	**1899**	**187**	**0**	**122**	**65**	**43**	**0**	**26**	**17**	**1,0000**	**0,1436**	**0,1436**	**1,0000**	**0,1792**	**0,1792**	**0,5980**
AT	6423	546	4866	1011	27	1	25	1	6	0	3	3	0,7241	0,0420	0,1095	1,0000	0,1579	0,0539	0,0149
CA	3745	302	2977	466	11	0	9	2	8	0	6	2	1,0000	1,0000	0,6374	1,0000	0,6709	0,2627	1,0000
CG	3745	255	3098	392	18	1	17	0	1	0	1	0	1,0000	0,3427	0,2471	1,0000	1,0000	1,0000	1,0000
CT	**3745**	**249**	**2358**	**1138**	**67**	**7**	**30**	**30**	**159**	**0**	**97**	**62**	**0,2141**	**0,0031**	**0,0155**	**0,0000**	**0,6160**	**0,0279**	**0,1676**
GA	**4230**	**156**	**3418**	**656**	**32**	**2**	**24**	**6**	**35**	**0**	**27**	**8**	**0,3339**	**0,3738**	**0,6225**	**0,6383**	**0,5246**	**0,2399**	**1,0000**
GC	4230	0	4042	188	5	0	5	0	1	0	1	0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
GT	4230	416	3626	188	24	3	20	1	37	0	33	4	0,7258	0,7678	1,0000	0,0455	0,8122	0,0826	0,6439
TA	6890	709	4422	1759	32	6	19	7	14	1	9	4	0,1362	0,5827	0,8390	1,0000	1,0000	0,7632	1,0000
TC	**6890**	**0**	**3428**	**3462**	**187**	**0**	**90**	**97**	**41**	**0**	**9**	**32**	**1,0000**	**0,7111**	**0,7111**	**1,0000**	**0,0004**	**0,0004**	**0,0028**
TG	6890	409	4891	1590	10	1	7	2	5	1	4	0	0,4583	1,0000	1,0000	0,2641	1,0000	0,5958	1,0000

Stop/Missense/Synonimous Frequencies per SNV type										
group	Stop	Miss	Syn	Stop	Miss	Syn	Stop	Miss	Syn
AC	0,0%	82,0%	18,0%	0,0%	93,3%	6,7%	0,0%	88,9%	11,1%
AG	**0,0%**	**70,4%**	**29,6%**	**0,0%**	**65,2%**	**34,8%**	**0,0%**	**60,5%**	**39,5%**
AT	8,5%	75,8%	15,7%	3,7%	92,6%	3,7%	0,0%	50,0%	50,0%
CA	8,1%	79,5%	12,4%	0,0%	81,8%	18,2%	0,0%	75,0%	25,0%
CG	6,8%	82,7%	10,5%	5,6%	94,4%	0,0%	0,0%	100,0%	0,0%
CT	**6,6%**	**63,0%**	**30,4%**	**10,4%**	**44,8%**	**44,8%**	**0,0%**	**61,0%**	**39,0%**
GA	**3,7%**	**80,8%**	**15,5%**	**6,3%**	**75,0%**	**18,8%**	**0,0%**	**77,1%**	**22,9%**
GC	0,0%	95,6%	4,4%	0,0%	100,0%	0,0%	0,0%	100,0%	0,0%
GT	9,8%	85,7%	4,4%	12,5%	83,3%	4,2%	0,0%	89,2%	10,8%
TA	10,3%	64,2%	25,5%	18,8%	59,4%	21,9%	7,1%	64,3%	28,6%
TC	**0,0%**	**49,8%**	**50,2%**	0,0%	**48,1%**	**51,9%**	**0,0%**	**22,0%**	**78,0%**
TG	5,9%	71,0%	23,1%	10,0%	70,0%	20,0%	20,0%	80,0%	0,0%

RNA/DNA vs Reference Contingency tables

C_i	$C_{total}-C_i$
CR_i	$CR_{total}-CR_i$

Where, for each outcome (i), C is the count of SNVs on transcriptomes or genomes, and CR is the count of potential SNVs on the viral reference genome.

genome vs RNA Contingency tables

$Cg_{missense}$	$Cg_{synonimous}$
$Cr_{missense}$	$Cr_{synonimous}$

Where Cg and Cr are, respectively, the count of SNVs on genomes and transcriptomes; nonsense SNVs has not been considered in this analysis.

Table 2: Missense/Nonsense/Synonymous mutations in SARS-CoV-2 transcriptomic and genomic data.

Last, this analysis is a first step in understanding the involvement of RNA editing in viral replication, and it could lead to clinically relevant outcomes: (i) If these enzymes are relevant in the host response to coronavirus infection, a deletion polymorphism quite common in the Chinese population, encompassing the end of APOBE*C3A* and most of *APOBEC3B* (146, 147), could play a role in the spread of the infection. (ii) Because RNA editing and selection act orthogonally in the evolution of the viruses, comparing genomic sites that are edited with those that are mutated could lead to the selection of viral regions potentially exploitable for therapeutic uses.

Section 2: Analysis of copy number variations from cell-free DNA of lung cancer patients via Nanopore sequencing

Introduction

Copy number variations and their impact on human diseases

Copy number variations (CNVs) are structural variants involving genomic segments of more than 1kb in length, which are represented in a variable number of copies compared to the normal ploidy of the organism (148).

For years, the weight of CNVs on human genome has been underestimated, also due to a lack of an adequate technology for their study. With the advent of large population studies, supported by more accessible and refined experimental approaches, CNVs have been recognised as a big contributor to inter-individual variation in the genomes of healthy individuals, along with single nucleotide polymorphisms (148, 149).

However, the presence of CNVs can influence the phenotype of the cells by altering the expression of the genes affected by the CNV or located nearby its boundaries (probably via alteration of adjacent regulatory sequences (148, 150, 151), and by generating to fusion-genes and, consequently, production of aberrant proteins (152).

It is hence not surprising that CNVs have been associated with a variety of diseases classified as 'genomic disorders'. Unlike mutation-driven pathologies which usually depend on variations on single genes, CNVs often involve large genomic regions affecting a set of genes as, for example, in Prader-Willi syndrome (15q11-q13 deletion) and Williams-Beuren syndrome (7q11.23 deletion) (153, 154).

However, this is not always the case: Smith-Magenis syndrome is caused by a deletion in chromosome 17p11.2; despite its size (on average 3.7 Mb) is variable among patients, a common 1.5 Mb portion has been identified. This "critical" portion includes the retinoic acid induced 1 gene, which is considered the main responsible of the pathologic phenotype (149, 155).

Germline inherited and de-novo CNVs have been associated with a wide spectrum of human diseases including:

- Infectious and autoimmune diseases: asthma, Chron's disease, HIV infections, systemic lupus erythematosus and anti-neutrophil cytoplasmic antibody-associated vasculitis (148, 156-168).
- Nervous system diseases: autism, schizophrenia, epilepsy, Parkinson's disease, amyotrophic lateral sclerosis and autosomal dominant Alzheimer's disease (169-177).
- Metabolic and cardiovascular diseases: familial hypercholesterolemia, atherosclerosis and coronary artery disease (178-180).
- Cancer (181-189).

This project is focused in particular in the detection of cancer related CNVs.

Cancer development is a multistep process characterized by the accumulation of genetic alterations eventually leading to the acquirement of the malignant phenotype (190). In contrast to most of the aforementioned pathologies, such alterations (including CNVs) can be both germline, representing a predisposing factor for the development of cancer, and somatic, contributing to the load of alterations necessary for the transformation (148, 191).

The relationship between CNVs and cancer development can be explained in part by the Kudson's "two hit" hypothesis (192): a homozygous deletion can lead to the loss of a tumor-suppressor gene, while a heterozygous one can be deleterious when the other allele is altered by an inactivating mutation or an additional deletion. On the other hand, amplifications can lead to overexpression of oncogenes. Notably, germline CNVs are typically more abundant in individuals from cancer-prone families, in particular among TP53 mutant carriers, suggesting that CNVs are not always a contributing cause to cancer, but rather a consequence of genomic instability (149, 193, 194).

Specific CNVs have been associated with cancer types and outcome: for example, in prostate cancer, loss of 8p23.2 is associated advanced stage disease, and gain at 11q13.1 is predictive predictive of post-operative recurrence; while heritable CNV at chromosome 1q21.1 is associated with neuroblastoma (189).

The recurrence of CNVs is not limited to single genes, but also to entire pathways such as the ERBB2, EGFR and PI3K pathways, which have been reported to be enriched in CNVs and SNVs in both breast and colorectal cancer (188).

Importance of cancer monitoring

During cancer development, malignant cells gain specific genotypic, phenotypic and epigenetic features, making cancer one the most heterogeneous human diseases.

Even within the same cancer type, it is possible to identify an large number of molecular subtypes defined by gene/protein expression patterns and alterations of the genome. Such molecular heterogeneity often results in different aggressivity, invasivity, response to treatment and, consequently, overall outcome (195-197).

For example, triple negative breast cancer patients have typically worse outcomes compared to HER2, progesterone and estrogen receptor expressing ones (198); EGFR mutations confer resistance to Tyrosine Kinase Inhibitor therapies in lung cancer patients (199); AR amplifications are linked to the development of castration resistance prostate cancers (200); and MGMT promoter methylation is an important prognostic biomarker, influencing the response to radio therapy in glioblastoma patients (201).

The goal of the so-called "precision oncology" is to define personalized treatment strategies based on cancer molecular features, aiming at maximizing the efficacy against a specific subtype. In this context, it is pivotal to accurately detect biomarkers for a proper (correct) subtype identification and patient stratification (195, 196).

Typically, bioptic samples or surgical resections are necessary for biomarker investigation. Tissue sections are typically used for techniques which are based on eye inspection such as: immunohistochemistry for protein expression, in-situ hybridization or RNA-scope for gene expression and fluorescent in-situ hybridization for structural variations (SV) detection. In addition, DNA and RNA can be extracted from tissue samples for gene expression, DNA methylation, mutation and copy number variation analyses (199, 202-210).

However, a significant limitation of tissue sampling is that it fails to comprehensively capture intra-tumoral heterogeneity. Indeed, cancer heterogeneity is not limited to inter-patient molecular diversity and a tumoral mass is often composed by subclones carrying different features. Hence, the portion of mass which is sampled may not fully represent the entire tumoral bulk (sometimes not even the major clone), with an high risk of missing clinically relevant alterations. Moreover, collection of tissue samples is usually invasive, requires trained medical staff and can be harmful for the patient. (210, 211).

Notably, cancer is an extremely dynamic disease: malignant cells are constantly under selective pressure, competing for nutrients against other cells or escaping human defense mechanisms, be them physiological (immune system) or artificial (drugs and treatments) (212). Consequently, the evolutionary path of each tumor can take different directions due to such pressure. It is hence important to monitor cancer development at multiple timepoints to closely follow its evolution, aiming at driving clinical decisions during the entire patient's history (195, 196). Unfortunately, the risks and invasiveness of conventional biopsy make it unsuitable for repeated sampling (213).

Liquid biopsy

A valid and non-invasive alternative to tissue sampling is represented by liquid biopsy. The principle behind liquid biopsy is that tumour masses shed cellular material into the bloodstream, urine or stool. It is therefore possible to analyse the blood to investigate tumor-related analytes to obtain information about the characteristics of the tumor (214). This concept is definitely not new as many protein-based serum biomarkers have proven useful for cancer diagnosis (215, 216). The most emblematic example is prostate specific antigen which is currently the first-line screening biomarker for prostate cancer early detection (216).

The recent emergence of cutting edge techniques with increased sensitivity and reliability allowed to extend blood-based analyses beyond proteic biomarkers.

It is currently possible to analyze tumor-derived nucleic acids (e.g. non coding RNAs, DNA) (213), vesicles (e.g. exosomes) (217) and circulating tumor cells (218) from blood samples.

In particular, I'm focused on the study of circulating cell-free DNA (cfDNA) which is extracellular DNA released into the bloodstream during cell death. CfDNA is extracted from plasma (less frequently from serum) obtained via blood centrifugation. Since we are interested in extracellular DNA, the goal of centrifugation is to remove intact blood cells whose DNA is non-informative and would reduce the sensitivity of the approach (219).

In healthy individuals, cfDNA belongs mainly from myeloid and lymphoid apoptotic cells due to the physiological turnover of hematopoietic cells (with minimal contributions from other tissues) (220, 221) while, in cancer patients, a fraction of the total cfDNA, termed circulating tumor DNA (ctDNA), comes from neoplastic lesions (222). CtDNA is very informative for the study of oncological pathologies as it harbours tumor-specific genetic alteration that reflects the genomic status of the malignant cell of origin (223-227).

However, several technical aspects make the study of cfDNA extremely challenging:

CfDNA concentration is very low and typically higher in advanced cancer patients rather than healthy subjects and low grade patients; this is one of the aspects that complicate cfDNA analysis, in particular for early-stage applications (214).

Moreover, the percentage of ctDNA among the totality of cfDNA can be very low (0.01-60%) [15–18] and depends on different tumor features such as tumor volume, stage, vascularization, proliferation rate, and cell death rate (223, 228-231).

For these reasons, cfDNA study requires highly sensitive techniques compatible with very low input DNA.

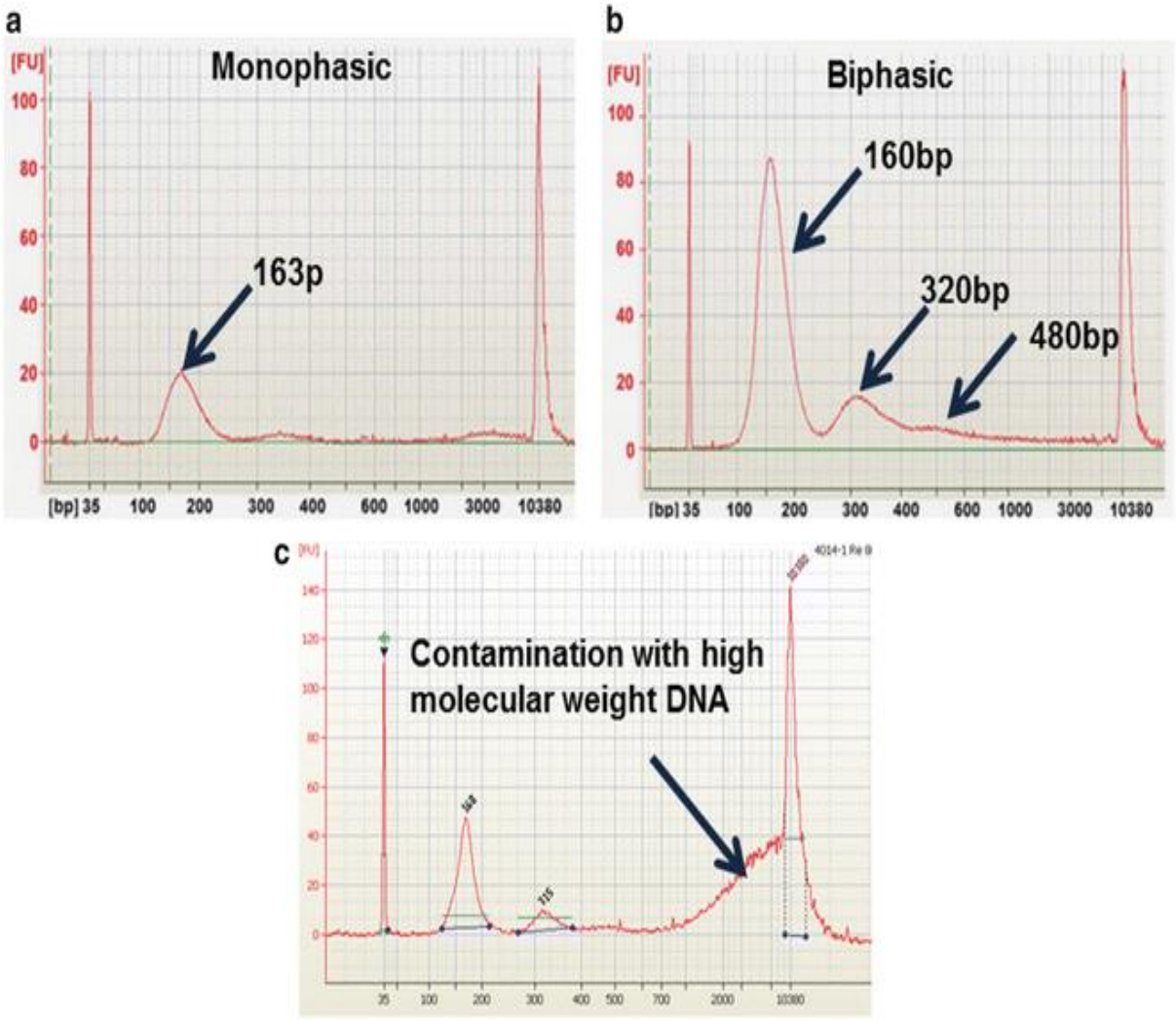

Figure 13: Typical fragmentation patterns in cfDNA. (A) Monophasic pattern showing only the ~160-167 bp peak. (B) Biphasic pattern showing ~160-167, ~320 and ~480 bp peaks. (C) Example of contamination by HMW DNA, likely due to blood cells lysis. Figure taken from (232).

CfDNA is also highly degraded with an enrichment in low molecular weight (LMW) DNA fragments: the typical cfDNA fragmentation profile is composed by a major peak at ~160-167 bp and two smaller peaks at ~320 and ~480 bp (not always detectable) (Figure 13) (221, 232). This particular pattern is due to nucleosome protection from degradation: due to cell death, DNA is degraded by DNAses, which cleave the filament mostly in linker regions between nucleosomes, generating mainly ~167 bp fragments corresponding to chromatosomes length (nucleosome + linker histone). Less frequent

scenarios, in which 2-3 chromatosomes are still intact, result in the weak peaks of longer DNA fragments (e.g. 320/480 bp) mentioned previously (221, 232).

The presence of high molecular weight (HMW) DNA is indicative of blood cell lysis, which can happen during sample transportation, blood withdrawal and centrifugation, or of an incorrect plasma collection, in which blood cells are accidentally resuspended by pipetting. HMW DNA is a dangerous contaminant for cfDNA, as it belongs to healthy blood cells and is hence uninformative (219).

Despite these technical challenges, liquid biopsy has some important advantages over conventional biopsy (211, 233), it is:

- Noninvasive: The entire procedure is completely unharmful and painless, without any complication that may arise with conventional biopsies.
- Simple: It is based on a simple blood withdrawal and it can be performed without particular training or instrumentation. This has also a positive impact on per-sample costs.
- Repeatable: This aspect is closely related to the aforementioned simplicity and non-invasivity that make liquid biopsy highly repeatable (even on a month-basis), just like any other routinary blood-based test.
- Versatile: The feasibility of tissue sampling depends on tumor characteristics
- Versatile and comprehensive: The feasibility of tissue sampling depends on tumor characteristics and location, meaning that not all tumors may be succesfully sampled. Virtually any cancer cell releases DNA into the circulation (as soon as it is vascularized), minimizing the risk of sampling biases. In addition, liquid biopsy provides a comprehensive profile of tumor, which include alterations from all the subclones and even from different tumor/metastasis sites.

These aspects make liquid biopsy preferable rather than tissue sampling in many contexts; however, the latter cannot be easily replaced since it is indispensable for histological analyses and most of the well-established biomarkers employed in clinical practice are based on conventional biopsy, while cfDNA field is still in an embryonic stage. Hence, further studies should be performed to fully exploit the potential of liquid biopsy, which should be seen more as a complementary tool rather than a complete replacement of tissue sampling.

Molecular-based methods for the study of CNVs

As for almost any other molecular investigation, the methods for the study of CNVs can be divided in targeted approaches, which investigate alterations in a locus-specific manner, or "omic" approaches, which provide a genome wide picture of the mutational landscape.

Targeted techniques:

- Quantitative real-time PCR (qPCR): is the oldest and the simplest PCR-based technique for the study of CNVs. It is based on the use of fluorescence signals for a real-time quantification of the amplicons. Fluorescence levels are proportional to the amount of amplified target DNA and are detected during the reaction. Fluorescence is obtained via dsDNA intercalants (typically Sybr green), or target complementary probes whose fluorescence is activated by the amplification of the target (Taqman chemistry). A fluorescence threshold is defined by the user, and the number of amplification cycles (CTs) necessary to cross the chosen

threshold are used as a proxy of amplicon abundance: the lower the CT the higher the abundance. The abundance of the target is compared to the abundance of a calibrator: a reference gene, which is supposed to be present in a known number of copies, depending on the expected ploidy. The ratio between target and reference abundances is used to calculate the number of copies of the target. However, the need of a calibrator represents a possible source of bias, since it is possible that also the expected copy number of the reference gene is altered by a CNV, leading to false positives/negative results (234-237).

- Digital PCR: It is the successor of qPCR, the main difference is the presence of thousands of micro reaction environments: some of them would contain the target molecule (positive) while others do not (negative). After the amplification, the positive reactions are detected by fluorescence detected, and counted providing a binary result (yes/no) rather than a continuous value (as in traditional qPCR). The counts of positive and negative droplets are related to the target's concentration by a Poisson function, used to infer the target abundance. With this principle is possible to make an absolute quantification of the target without the use of a standard curve. However, for CNV the use of a reference gene as calibrator is still suggested. The nature of the reaction environment depends on the manufacturer, the most used versions involve micro reaction wells on a chip (Thermofisher) or droplets produced by emulsion (Biorad, digital droplet PCR, ddPCR). The main advantage over classical qPCR is the high reliability of the approach, even at very low concentrations, which make it ideal for liquid biopsy studies (236, 237).
- Multi Ligation-dependent Probe Amplification (MLPA): It is a multiplex PCR-based method involving target-specific sets of probes. Each set of probes is made

of a 5’ and a 3’ half-probes which hybridize to the target sequence. After the hybridization, the inner ends of each probe are adjacent and can be connected by a ligation reaction. Each probe contains a primer binding site for subsequent PCR amplification of the ligation product. The 3’ probes contain also a stuffer sequence of variable length with the aim of creating PCR amplicons of different length. The primers complementary to the 5’ probe binding sites are fluorescently labeled. The products of amplification are then analyzed with capillary electrophoresis which detects fluorescence intensity (as a proxy of amplicon abundance) and each set of probes is discriminated by migration time (as a proxy of amplicon length). Abundances are then normalized with the results obtained from a control sample to infer the number of copies of the targets (238). MLPA is a cost effective approach when analyzing small sets of targets (~40), while other PCR-based approaches are usually preferable when studying single targets; however, it typically requires higher amounts of high quality input DNA. Hence, the high fragmentation of cfDNA can be a big obstacle for the use of this technique. Despite these limitations, MLPA has been successfully exploited for CNV detection in a liquid biopsy context (239).

“Omic” techniques:

- Comparative Genomic Hybridization (CGH) arrays: A large number of spots containing target-specific probes are immobilized onto the glass surface of a chip. A library of fluorescently labelled DNA is prepared and deposited on the chip for probe hybridization. The library is composed of DNA coming from the sample of interest and a reference DNA (usually a pool of DNA from healthy individuals) marked with two different fluorophores. After the hybridization, the chip is

washed, and the fluorescence in every spot, proportional to the amount of bound DNA, is detected. CNVs are determined by detecting differences in the spots fluorescence levels between the sample of interest and the reference DNA. The main advantage of CGH arrays lies in their genome-wide nature, which allows the discovery of CNVs without any prior knowledge about their genomic position. On the other hand, the resolution of the array is usually lower than PCR-based methods and depends on the number of probes, allowing the detection of CNVs as small as 5-10kb. Once again, CGH arrays are not the ideal method for liquid biopsy-based analyses since an high fraction of tumor derived DNA (>50% or, ideally, 80–90% or higher) is required, which is definitely not the case for cfDNA (148, 240).

- Single-Nucleotide Polymorphism (SNP) Arrays: They employ the same array-based principle of CGH arrays. In this case, the information about the copy number status is a byproduct, since they are designed for the detection of SNPs.
  For each target a set of probes specific for different alleles is used, and the genotype is determined by measuring the allele-specific fluorescence. CNVs are determined based on the total measured intensities: large CNVs spanning multiple SNPs have intensity ratios patterns distinct from normal disomic
  regions. SNP arrays share the same limitations of CGH arrays; in addition, certain genomic loci are particularly difficult to analyze for SNP detection and are therefore often removed from the array, with a consequent lack of data for CNV detection in that area (148, 240).
- Deep sequencing: It represents the most advanced approach for genome-wide CNV detection and, since it is the focus of this section of the thesis, it will be thoroughly addressed in the next paragraph.

Sequencing-based CNV analysis from cfDNA

Sequencing-based CNV analysis takes advantage of DNA Whole Genome Sequencing (WGS) data, typically produced with SGS experiments. Most of the software for bioinformatic analysis is based on read count (RC), which is a proxy of genomic abundance: compared to diploid regions, the relative amount of reads produced during the sequencing would be higher for amplified regions and lower for deleted regions. The genome is hence divided in windows (or "bins") of fixed length (typically 100kb - 2Mb), and the RC is obtained by counting the number of reads mapping to each bin. The RC is then normalized taking in account for mappability and GC content, which are two of the main sources of noise for this kind of analysis. Subsequently the RC for each bin is expressed as log2ratio values under the following assumption:

$$Log2ratio = log2(RC/EXP)$$

Where EXP is the expected RC for diploid regions calculated as following:

$$EXP = \frac{N * L}{G}$$

Where N is the total number of generated reads, L the length of the region of interest (for example a bin) and G is the length of the genome.

Lastly, segmentation algorithms are applied to noisy log2ratio to define the boundaries of CNVs. The noise of log2ratio is inversely proportional to the sequencing depth which can be tuned by the experimentator, depending on the resolution needed (241-243).

Notably, the total number of reads needed for this approach is relatively limited, with accurate detection of CNVs even with a very low coverage (<1X) sequencing experiment, usually termed “low-pass sequencing”, or “Shallow Whole Genome Sequencing” (SWGS). The advantage of SWGS over high-coverage WGS is the small throughput requirement that allows the use of entry-level sequencers and reduce the per-sample costs (244, 245). On the other hand, with high-coverage WGS is possible to precisely detect breakpoint location employing discordant pairs of reads and split reads, and to reduce the bin size in order to detect smaller CNVs (241).

During the years, a multitude of tools for CNV analysis from SGS reads has been released; they often differ for the segmentation algorithm and normalization approaches, but the main principle is roughly the same (241).

Currently, low-input WGS protocols are available that make this approach compatible with low cfDNA fragmentation; also, the required tumoral DNA fraction is substantially lower (~10%) than array-based assays (245, 246).

SWGS is indeed a powerful technique for liquid biopsy applications, but the need for expensive SGS instruments (typically Illumina sequencers) is often an obstacle for smaller laboratories. Also, Illumina cfDNA and low-input protocols usually involve several PCR steps during library preparation, which can introduce bias and reduce the performance of CNV analysis (247).

Third generation sequencing

Oxford Nanopore Technologies (ONT) has recently released MinION: a fast and extremely inexpensive third generation sequencer. Nanopore technology is based on an array of nano-scale proteic holes fixed on an dielectric polymer membrane. The passage

of a single nucleic acid filament through a pore produces an electric signal which depends on its sequence (248). This electrical current signal (a.k.a. the 'squiggle' due to its appearance when plotted) is the raw data gathered by an ONT sequencer. Basecalling for ONT devices is the process of translating this raw signal into a DNA sequence. It is not a trivial task as the electrical signals come from single molecules, making for noisy and stochastic data. Furthermore, the electrical resistance of a pore is determined by the bases present within multiple nucleotides that reside in the pore's narrowest point (~5 nucleotides for the R9.4 pore), yielding a large number of possible states: $2^5 = 1024$ for a standard four-base model (249).

Most of the current basecallers divide the raw current signal into discrete blocks, which are called events. After event-detection, each event is decoded into a most-likely set of bases. In the ideal case, each consecutive event should differ by one base. However, in practice, this is not the case because of the non-stable speed of the translocation.

Also, determining the correct length of the homopolymers is challenging. Both of these problems make deletions the dominant error of nanopore sequencing (250).

R9.4 is the most widely used version of the pore, characterized by a relatively high throughput but a significant error rate (~5-20%) (248); R10 pore it has recently been released with a longer barrel and dual reader head, enabling improved resolution of homopolymeric regions and improving the consensus accuracy of nanopore sequencing data (https://nanoporetech.com/).

Unlike SGS technology, there is no need of fragmentation and read length can reach several kilobases. For this reason, Nanopore is particularly indicated for non numeric structural variations, breakpoint detection, isoform quantification and fusion-transcript detection, for which long-read sequencing is the approach of choice (248, 251).

After the passage of a filament, the pore becomes available for the sequencing of a new molecule, and the electric signal produced is immediately stored and ready for analysis. This aspect is crucial as it allows the user to obtain sequencing results and perform real-time analyses while the instrument is still running (248, 252).

No PCR amplification is needed for library preparation, reducing PCR-related biases (243, 247) and preserving base modifications . This feature allows the detection of DNA methylation without conversion (e.g. bisulfite treatment) via direct sequencing: modified bases produce specific shifts in the electric signal that can be used to discriminate them (253).

The main drawback of Nanopore technology is the high error rate which complicates accurate SNVs detection, in particular when dealing with somatic variants (248). The high error rate is compensated by the length of the read produced: the longer the read the easier the alignment, even in presence of a high number of mismatches (254).

The aim of my project is to exploit Nanopore sequencing for the study of CNVs from cfDNA of cancer patients; the use of this technology offers many advantages over classical SGS (248, 255):

- PCR-free workflow: lack of PCR amplification in Nanopore workflow prevents SGS-typical biases which would otherwise hamper CNV detection.
- Real-Time sequencing: the parallel nature of SGS allows the user to analyse the results only at the end of the run, which can last several hours. With Nanopore-seq, it is possible to analyse the results in real-time during the run, allowing the user to detect CNV as soon as the necessary amount of reads is produced. Nanopore flow cells can be washed and re-used for a new library; hence, after the generation of a satisfactory number of reads, it is possible to stop the run and exploit any residual sequencing power of the flow cell for other runs.

- Scalability: typically, SGS sequencing costs are competitive if multiplexing several samples in the same flow cell; on the other hand, the cost-effectiveness decreases when reducing the number of pooled samples. Nanopore Flongle flow cells have a reduced number of pores compared to regular Nanopore flow cells; their reduced cost can drastically increase the cost-effectiveness of small-scale experiments.
- Minimal instrumentation costs: MinION is the entry level Nanopore sequencer, its cost is extremely low (~ 1,000 €) compared to other sequencers whose price is in the order of tens of thousands of euros. Reduced instrumentation costs makes this technology accessible to most of the laboratories which are otherwise forced to resort to sequencing companies, or to access shared sequencers (not always) available in their institution, leading often to long queues and delays.

Unfortunately, Nanopore technology is optimized for long read sequencing, hence it is not ideal for sequencing of short cfDNA fragments. Indeed, standard Nanopore protocols involve several clean-up steps that are designed to preferentially retain long DNA fragments with a consequent loss of short fragments. This is probably the reason why, previous attempts to sequence cfDNA samples produced a very limited amount of throughput (256).

For this reason, it is necessary to develop customized workflows to adapt Nanopore-seq to plasma cfDNA, in order to exploit its potential also for liquid biopsy applications.

Rationale

The aim of the study is to assess the feasibility of Nanopore sequencing of cfDNA by modifying standard protocols, and to compare its performance in CNVs detection with state of the art approaches, namely Illumina sequencing.

A pre-print, including the following content, entitled “Nanopore sequencing from liquid biopsy: analysis of copy number variations from cell-free DNA of lung cancer patients” https://doi.org/10.1101/2020.06.22.165555, is available at www.biorxiv.org.

Methods

The aim of this project is to set-up a workflow for the identification of whole genome CNVs from cfDNA using Nanopore technology. The approach is based on shallow whole genome sequencing, which is a read-count based approach that allows detection of genome-wide CNVs from reads produced through a low-coverage (< 1x) whole genome sequencing experiment (257).

Since Nanopore library preparation protocols are designed to enrich long DNA fragments, we have modified them in order to retain small cfDNA fragments. With our custom protocols, we sequenced cfDNA from 6 cancer patients and 5 healthy subjects, in both singleplex and multiplex runs (S1, M1 and M2, Table 3). To validate our workflow, 4 patients were sequenced also with SGS platforms (Illumina) and results were compared.

Case Series			Run Stats						Mapping Stats				Illumina Sequencing		
Sample	Group	Sex	Run	Multiplexing	Input DNA (ng)	Raw Reads	Total Run Throughput (Reads)	Relative Throughput (%)	Mapped reads BWA	mapped reads BWA (%)	Mapped reads Minimap2	Mapped Reads Minimap2 (%)	Raw Reads	Mapped Reads	Mapped Reads (%)
19_744	Tumor	M	M1	Yes	30	15240945	31582051	48,26	15162985	99,5	13677899	89,7	17259281	17237658	99,9
HM1	Healthy	M	M1	Yes	30	3683399	31582051	11,66	3642480	98,9	3284644	89,2	NA	NA	NA
HM2	Healthy	M	M1	Yes	15	3187147	31582051	10,09	3153485	98,9	2864136	89,9	NA	NA	NA
HM3	Healthy	M	M1	Yes	30	6486314	31582051	20,54	6417802	98,9	5904030	91,0	NA	NA	NA
HF1	Healthy	F	M1	Yes	60	2984246	31582051	9,45	2938942	98,5	2652708	88,9	NA	NA	NA
18_1130	Tumor	F	M2	Yes	30	8221101	19610131	41,92	8090242	98,4	6560697	79,8	23960716	23907788	99,8
19_1231	Tumor	F	M2	Yes	30	4068177	19610131	20,75	4011152	98,6	3277098	80,6	22846324	22806737	99,8
19_560	Tumor	M	M2	Yes	30	2953200	19610131	15,06	2920132	98,9	2477759	83,9	NA	NA	NA
19_924	Tumor	M	M2	Yes	30	4031897	19610131	20,56	3977333	98,6	3265747	81,0	22325079	22295913	99,9
HF2	Healthy	F	M2	Yes	30	335756	19610131	1,71	327021	97,4	273625	81,5	NA	NA	NA
19_326	Tumor	M	S1	No	25	14338633	14338633	100,00	13921937	97,1	12348607	86,1	NA	NA	NA

Table 3: Case series and run statistics.

Sample collection and cfDNA isolation

Blood from 5 unrelated healthy donors and 6 unrelated metastatic Non Small Cell Lung Cancer patients was collected in EDTA vacuum tubes. Blood samples were centrifuged at 1600g x 10'', and plasma was carefully collected with a pipet without disturbing sedimented blood cells.

cfDNA was extracted from 4ml of plasma using QIAamp Circulating Nucleic Acid Kit (QIAGEN, 55114), it was quantified via Qubit Fluorometer (Thermo Fisher Scientific, dsDNA HS assay kit, Q32851), and its fragmentation pattern was obtained via Agilent 2100 Bioanalyzer (Agilent, High Sensitivity DNA kit, 5067-4626). Extracted cfDNA was stored at -80° C.

Nanopore library preparation and analysis

For library preparation, the EXP-NBD104 and SQK-LSK109 protocols were used: the bead/sample ratio of AMPure XP beads (Beckman Coulter, A63880) was increased to 1.8x in all clean-up steps.

All the other steps were performed following the manufacturer's instructions.

The SQK-LSK109 protocol was used for the run S1. In the case of the multiplex runs M1 and M2, 25ul of each barcoded sample were pooled together before adapter ligation. The pool was then cleaned-up using 2.5X AMPure XP beads.

S1, M1 and M2 runs were performed using FLO-MIN106 (R9.4) flow cells on a GridION sequencer. FASTQ files were generated using real-time high-accuracy basecalling during the run with the MinKNOW software (version 18.12.9); guppy (version 1.8.10) was used for the actual basecalling with SQK-LSK109 and FLO-MIN106 settings. Porechop

(https://github.com/rrwick/Porechop) was used to de-multiplex FASTQ files of multiplex runs (M1, M2), and to trim adapters of all the runs.

Minimap2 (with *-ax map-ont* flags) (254) and BWA mem (with *-x ont2d* flags) (112) were used to align raw reads, using the human_g1k_v37_decoy as reference genome.

The CIGAR field of aligned BAMs was used to determine fragment length of sequenced cfDNA (Figure 14).

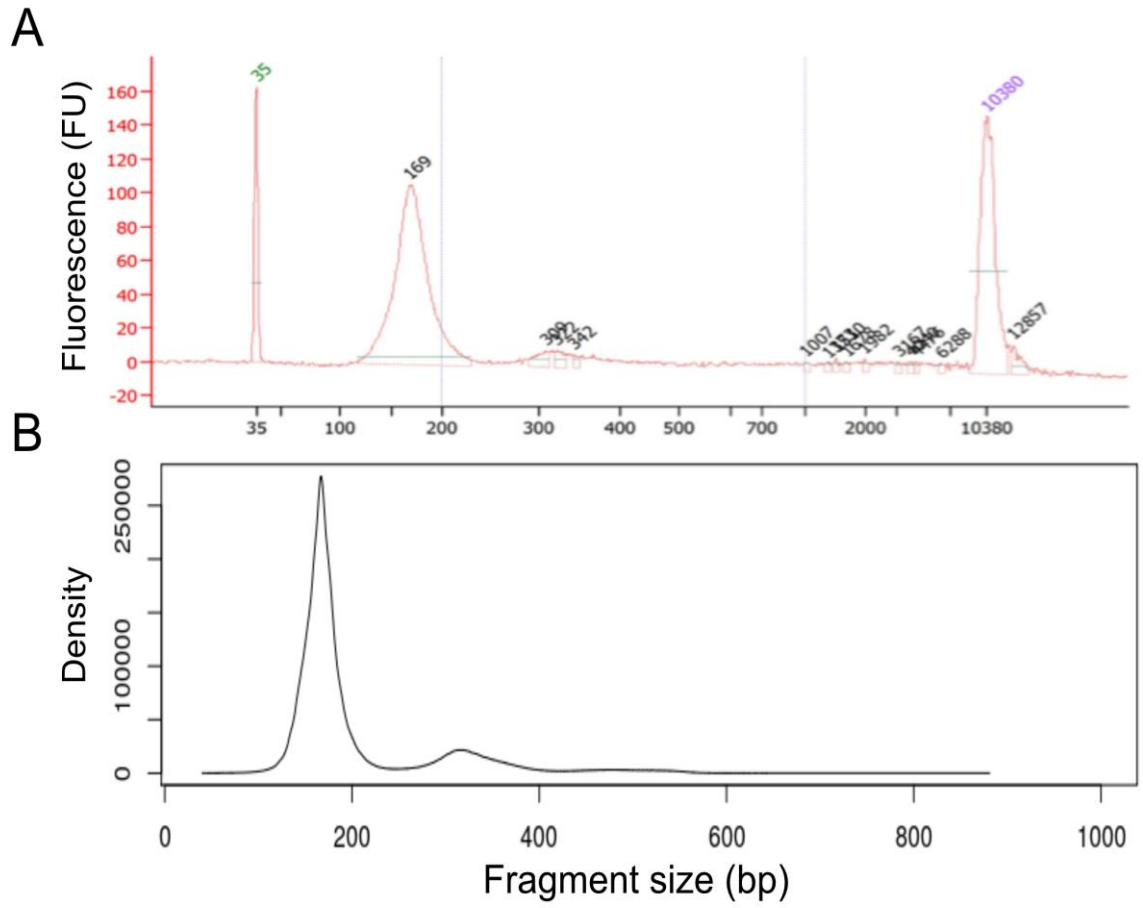

Figure 14: Fragment size distribution. Fragment size distribution estimated via Bioanalyzer (A), and from Nanopore reads (B).

NanoGLADIATOR was used to generate molecular karyotypes of BWA aligned BAMs with a bin size of 100kb (243).

For “paired” mode analysis, HF1 was used as a control for female patients, and BAMs from HM2 and HM3 were merged and used as control (Healthy_Males_Pool, HMP) for male patients.

Additional details on patients features, library preparation and run statistics are summarized in Table 3.

Illumina library preparation and analysis

Illumina libraries for samples 19_924, 19_744, 19_1231 and 18_1130 were prepared from 15ng of input DNA (from the same cfDNA extraction of the DNA used for Nanopore library preparation), using Ovation Ultralow V2 DNA-seq Library Preparation Kit (NUGEN, 0344NB-A01), sequencing runs (150bp, paired end) were performed on a NovaSeq 6000 sequencer (Illumina).

Only R1 reads were used for CNV analysis treating them as the product of a single end sequencing experiment, in order to simplify subsequent steps such as subsampling and comparison with Nanopore results. This strategy doesn’t introduce any methodological bias, since Illumina single-end and paired-end CNV results are highly correlated (Table 4).

FASTQ files were aligned with BWA mem using human_g1k_v37_decoy as reference genome.

XCAVATOR was used to generate molecular karyotypes of BWA aligned BAMs with a bin size of 100kb (242).

Case	*Single-end vs Paired-end*								*2M reads vs Full depth (single-end)*							
	Whole genome				**Copy number altered regions**				**Whole genome**				**Copy number altered regions**			
	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)
19_744	99,69	0,00	0,97	0,99	99,34	0,00	0,97	1,00	93,83	0,00	1,08	0,95	88,65	0,00	1,08	0,95
18_1130	99,89	0,00	1,00	1,00	99,77	0,00	1,00	1,00	94,37	0,02	0,99	0,96	92,58	0,03	0,98	0,96
19_924	99,61	0,00	1,00	0,99	98,53	0,00	1,00	0,99	94,68	0,00	0,99	0,91	94,21	0,00	1,00	0,86
19_1231	99,80	0,00	1,00	1,00	99,74	0,00	1,00	1,00	97,28	0,00	0,98	0,99	96,64	0,00	0,98	0,98

Table 4: Correlation of Illumina results. Paired-end Vs single-end, and subsampled BAMs (2M reads) Vs full BAMs

Segmentation comparison

Custom R scripts were used to compare segmentation results:

When comparing two experiments, the "segment mean" value of each of the 100kb bins was correlated (corr.test function, R base package, method="*spearman*").

To determine the percentage of genomic positions with concordant copy number status, we considered two bins as "concordant" if their segment mean differs by ±0.08.

Chromosome Y bins were ignored when analysing female patients.

When comparing Illumina and Nanopore results, even if the bin size used was the same, the starting positions of the bins slightly differs among the two pipelines; an XCAVATOR bin is considered corresponding to a NanoGLADIATOR bin if its starting position falls between the starting and the end position of the NanoGLADIATOR bin. Only NanoGLADIATOR bins for which it was possible to identify a corresponding XCAVATOR bin were considered for subsequent analysis.

In total, 26867 were used for segmentation comparison.

The original number of bins for respectively Nanopore and Illumina was 26927 and 28452 (0,2% and 5.6% bin loss).

Results

Sequencing yield and quality control

With our custom protocols, we obtained 14,338,633, 19,610,131, and 31,582,051 raw reads from the S1, M1 and M2 runs, respectively: a remarkably higher throughput than previously reported (256) (Table 3). Notably, the per-sample throughput was highly variable, even if the amount of input DNA was constant for most of the samples (30ng). Indeed, for sample HF2,

the throughput obtained was insufficient. To assess the effects of input DNA on per-sample throughput, we performed library preparation of samples HM2, HM1 and HF1 with respectively 15, 30 and 60 ng of DNA; however, the amount of reads produced was very consistent among the three samples (~3M reads, Table 3), suggesting that input DNA has a low impact on the final throughput.

For the run M2, we quantified eluted DNA after each clean-up step via Qubit Fluorometer: Since DNA concentration highly correlates with read yield, differences in per-sample yields are likely attributable to a different efficiency of library preparation steps rather than amount of input DNA.

Nanopore protocols suggest pooling equimolar quantities of barcoded samples prior to adapter ligation to avoid differences in per-sample throughput. However, in order to avoid any waste of DNA and aiming at obtaining the maximum amount of reads from a single flow-cell, we loaded the entire barcoded sample for each patient, which may explain the observed variability.

Unexpectedly, the relative-throughput (sample reads/total run reads) of cancer patients is remarkably higher compared to healthy subjects (Table 3**).** Since there were no differences in input DNA, and per-sample throughput depends mainly on library preparation efficiency, it is possible that the presence of ctDNA positively affects library preparation efficiency; however, the biological aspects of this behaviour are not clear and should be further investigated.

A possible explanation is that cancer samples typically show a higher proportion of smaller fragments (258), which can be somehow enriched by our modified protocol.

Alignments were performed with both BWA and Minimap2. The average percentage of uniquely mapped reads was 98.5% and 85.6%, respectively (Table 3). Size distribution of the sequenced cfDNA fragments perfectly matches the fragmentation profile obtained with Agilent Bioanalyzer (Figure 14). Fragment size distributions of healthy and cancer patients

are comparable; notably, in cancer patients, a higher degradation in fragments belonging to the 332 bp peak is observed (Figure 15). While Minimap2 is usually recommended for alignment of long Nanopore reads, according to our results BWA is preferable for cfDNA-derived data, probably due to the shorter length of cfDNA fragments.

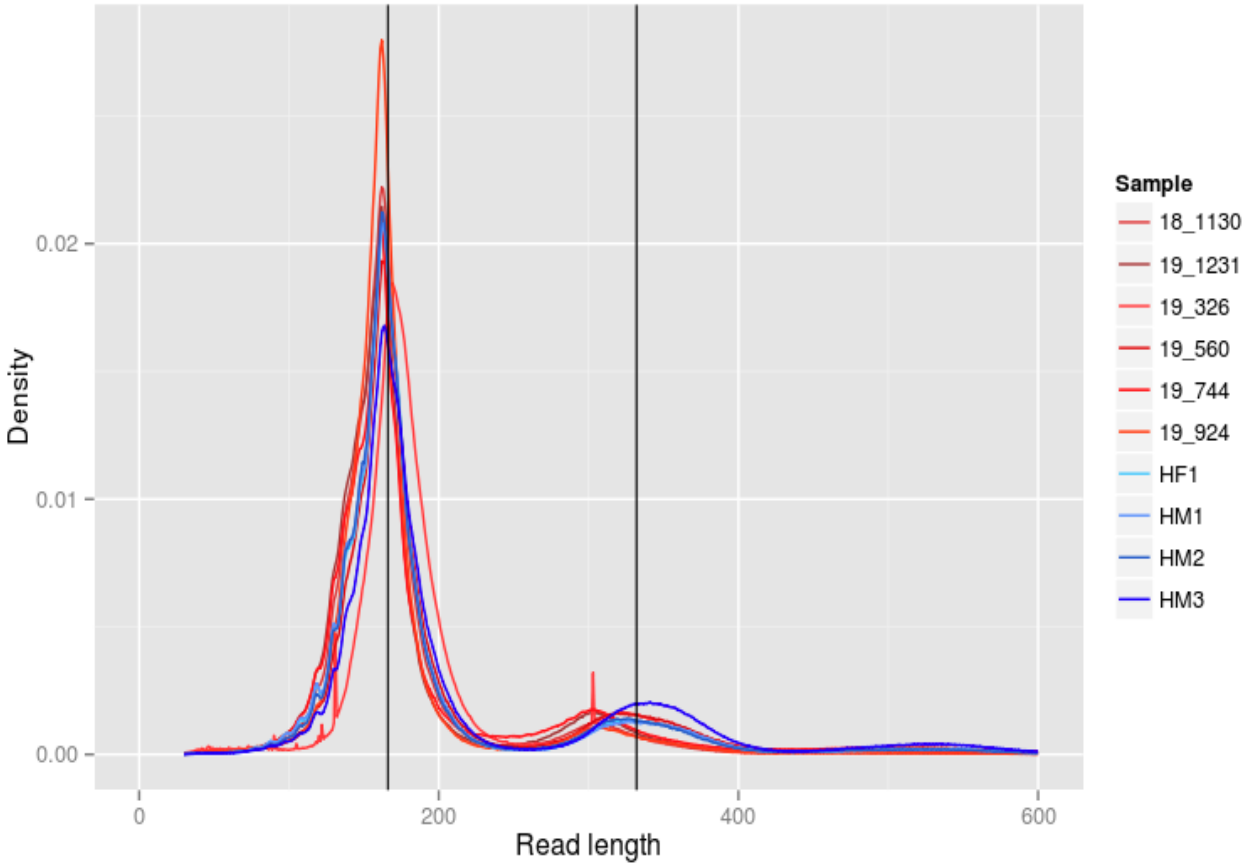

Figure 15: Read length distribution for healthy (blu scale) and cancer patients (red scale). Vertical lines highlight 166 and 332 bp sizes.

CNV profiling and artefact removal

Molecular karyotype of 10 out of 11 samples was successfully produced using NanoGLADIATOR (*"nocontrol"* mode), a recently developed tool for the identification of CNVs from read counts (reported as log2ratio) across multiple consecutive windows (bins) (243). BWA-aligned BAM files were analysed with a bin size of 100kbp, and CNVs were detected in all the tumoral samples (Figure 16).

Unexpected variations in read-count values were present also in samples from healthy donors (Figure 16**)**. Most of the variations observed in healthy donors are shared by at least 2 healthy subjects, suggesting that they may be errors introduced by the technique itself rather than patient-specific alterations (Figure 17). Even though it is possible these variations represent naturally occurring polymorphisms, this is unlikely: polymorphic variations should present a discrete number of copies (1,3 or 4 copies), which is not the case, as most of these variations have weak log2ratio.

These technical artefacts can be easily filtered out setting a threshold. On the other hand, some of these variations are very similar in terms of length and segment mean (roughly ±0.10) to those we observe in cancer samples and it could be difficult to discriminate real CNVs from these ones (Figure 16). Typically, these artefacts are present in regions containing a higher number similar sequences, e.g. the sexual chromosomes (Figure 16). Alignment of short reads in such genomic regions is typically challenging and presence of these artefacts is likely due to mapping issues (259). In order to minimize the number of artefacts, we used NanoGLADIATOR in *"paired*" mode, which generates segmentation results comparing test samples with a control sample. We tested this approach on healthy male subjects, using each sample as both case and control, in any possible combination.

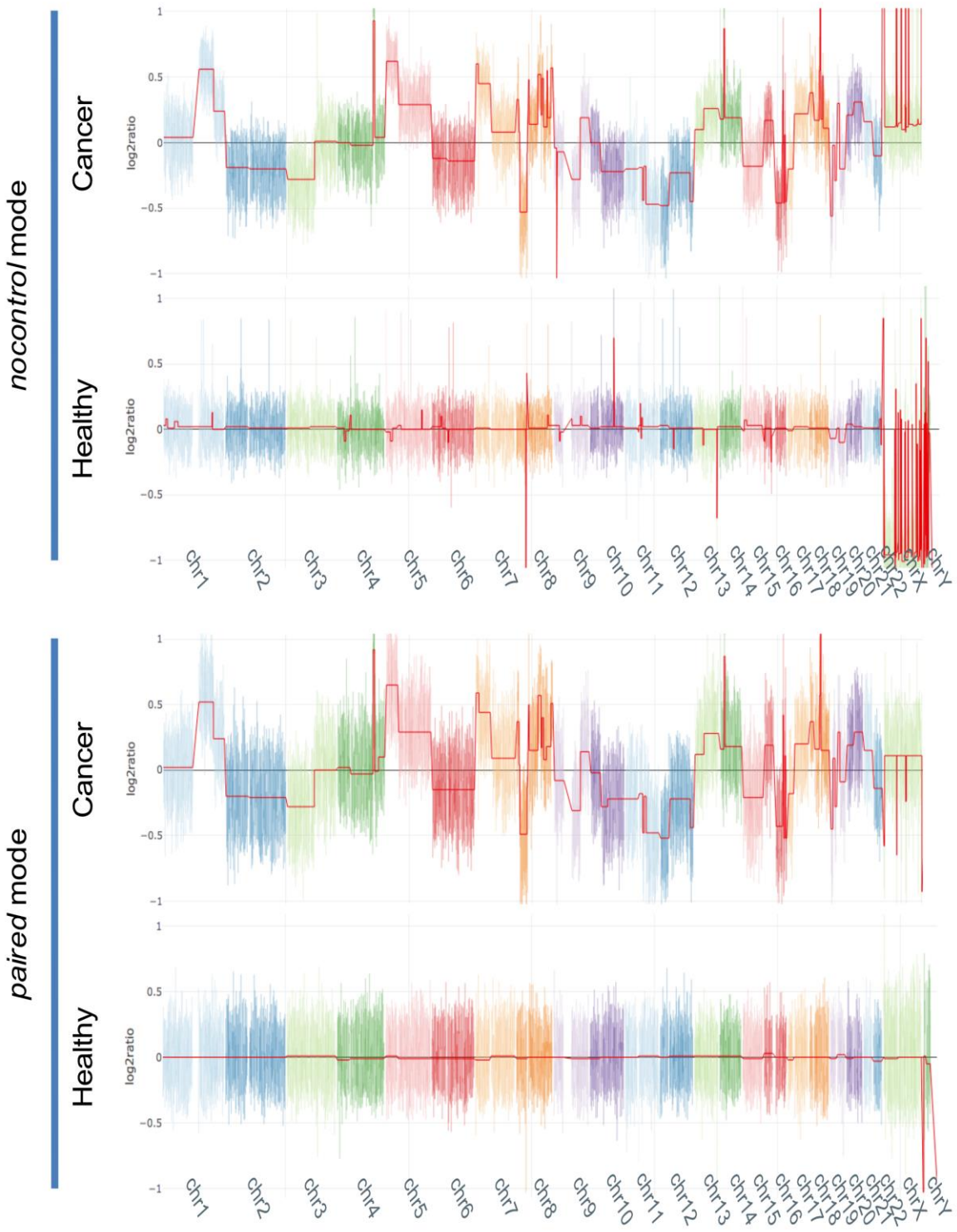

Figure 16. NanoGLADIATOR segmentation plots. Segmentation plots produced with NanoGLADIATOR for samples 19_1231 (cancer) and HM3 (healthy) in "nocontrol" mode (A), and "paired" mode (B). In "paired" mode, HF1 and HM2 were used as controls for respectively 19_1231 and HM3. The red line indicates the segment mean (log2ratio). Each color represents a different chromosome; chromosome Y for sample 19_1231 has been omitted.

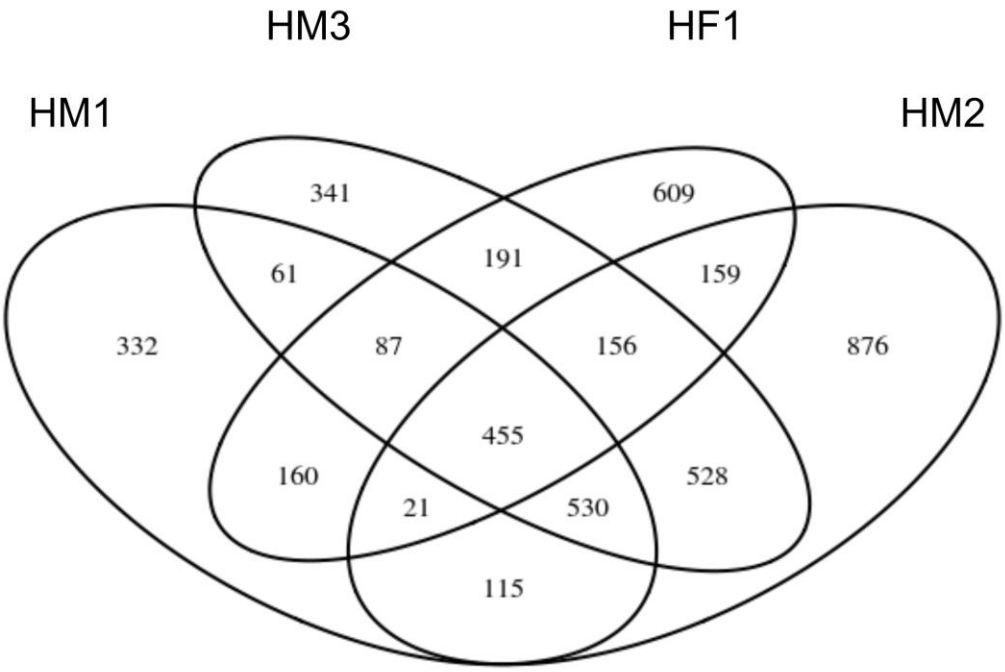

Figure 17. Technical artifacts in healthy samples. Venn diagram reporting recurring genomic bins with altered log2ratio in healthy samples.

Using this strategy, we were able to remove 82-100% of false positive bins in healthy males samples (segment mean threshold ≥ 0.04, or ≤ -0.04) (Table 1**,** Figure 16). Moreover, when HM1 is not used neither as a case nor a control, the number of false positive bins is reduced by 100% (Table 5), suggesting that this sample might be enriched in sample-specific artefacts. Hence, HM1 was not used as a control in subsequent "paired" analyses to avoid introduction of biases. HM2 and HM3 BAM files were merged and the resulting BAM (HMP) was used as control for male patients, while HF1 was used as control for female patients.

This approach doesn't negatively affect the performance of the analysis, as the number of copy-number altered bins is reduced by less than 5% in most of the tumoral samples and increases by 29% in sample 19_744; sample 19_560 is the only exception, with a reduction of roughly ~40% (Table 5). 19_560 shows the lowest number of altered bins

Case	Control	*Nocontrol* mode			*Paired* mode			(Altered bins (*paired*) - Altered bins (*nocontrol*)) / Altered bins (nocontrol)
		Altered bins	Altered segments	Segment mean standard deviation	Altered bins	Altered segments	Segment mean standard deviation	
19_744	HMP	15979	134	0,14	20566	111	0,14	0,29
18_1130	HF1	24775	161	0,25	24110	121	0,25	-0,03
19_924	HMP	14407	93	0,07	13746	59	0,07	-0,05
19_1231	HF1	23601	94	0,27	22775	76	0,27	-0,03
HM1	HM2	1761	46	0,03	317	6	0,01	-0,82
HM1	HM3	1761	46	0,03	182	4	0,02	-0,90
HM2	HM1	2840	48	0,03	317	6	0,01	-0,89
HM2	HM3	2840	48	0,03	0	0	0,01	-1,00
HM3	HM1	2349	69	0,03	182	4	0,02	-0,92
HM3	HM2	2349	69	0,03	0	0	0,01	-1,00
HF1	-	1838	31	0,03	-	-	-	
HF1	-	1838	31	0,03	-	-	-	
19_560	HMP	6571	77	0,05	3991	36	0,05	-0,39
19_326	HMP	24495	130	0,23	23075	114	0,22	-0,06

Table 5. Performance of NanoGLADIATOR pipeline in "nocontrol" and "paired" mode

and the lowest segment mean standard deviation (calculated on autosomes) (Table 5). A lack of clonal CNVs in the tumor, or a lower concentration of ctDNA fragments among the overall cfDNA population can explain these observations; it is therefore not surprising to observe an "healthy-like" genotype, with false positives representing a large part of the detected CNVs. Notably, this sample is also the one wit the lowest number of reads, but it is unlikely that this is affecting the results, as low coverage results highly correlate with full-depth results (see next paragraph).

Using NanoGLADIATOR in "paired" mode allows to set a very strict log2ratio threshold (±0.04) to discriminate technical artefacts from real CNV, drastically increasing the performance of the approach in terms of sensitivity/specificity (Figure 16**,** Table 5).

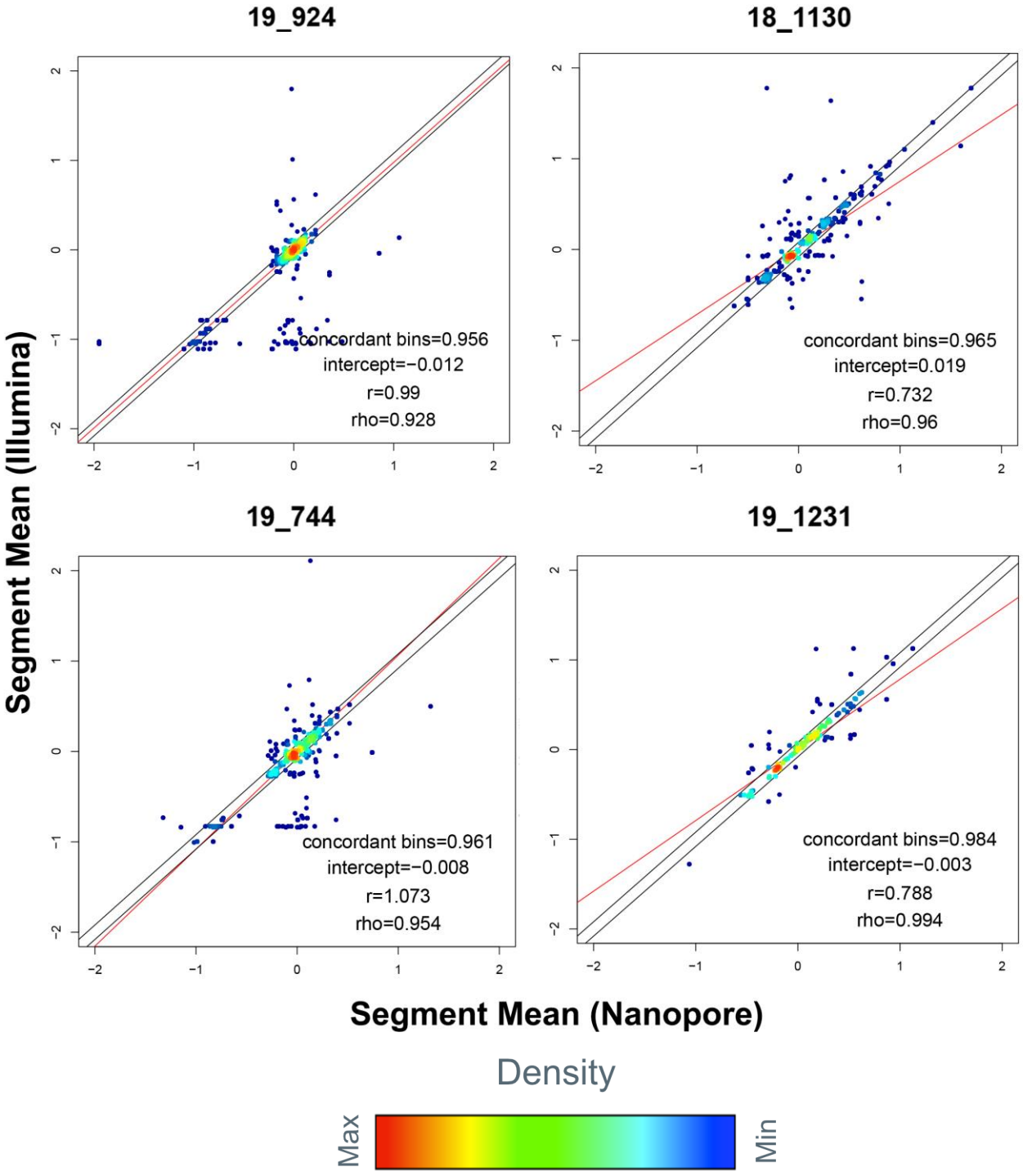

Figure 18: Comparison of segmentation results of cancer patients with Nanopore and Illumina: Correlation of Nanopore and Illumina segment mean values. Each genomic bin is represented as a dot, colours indicate dot density. Regression lines are shown in red. Black lines indicate the thresholds for concordant bins.

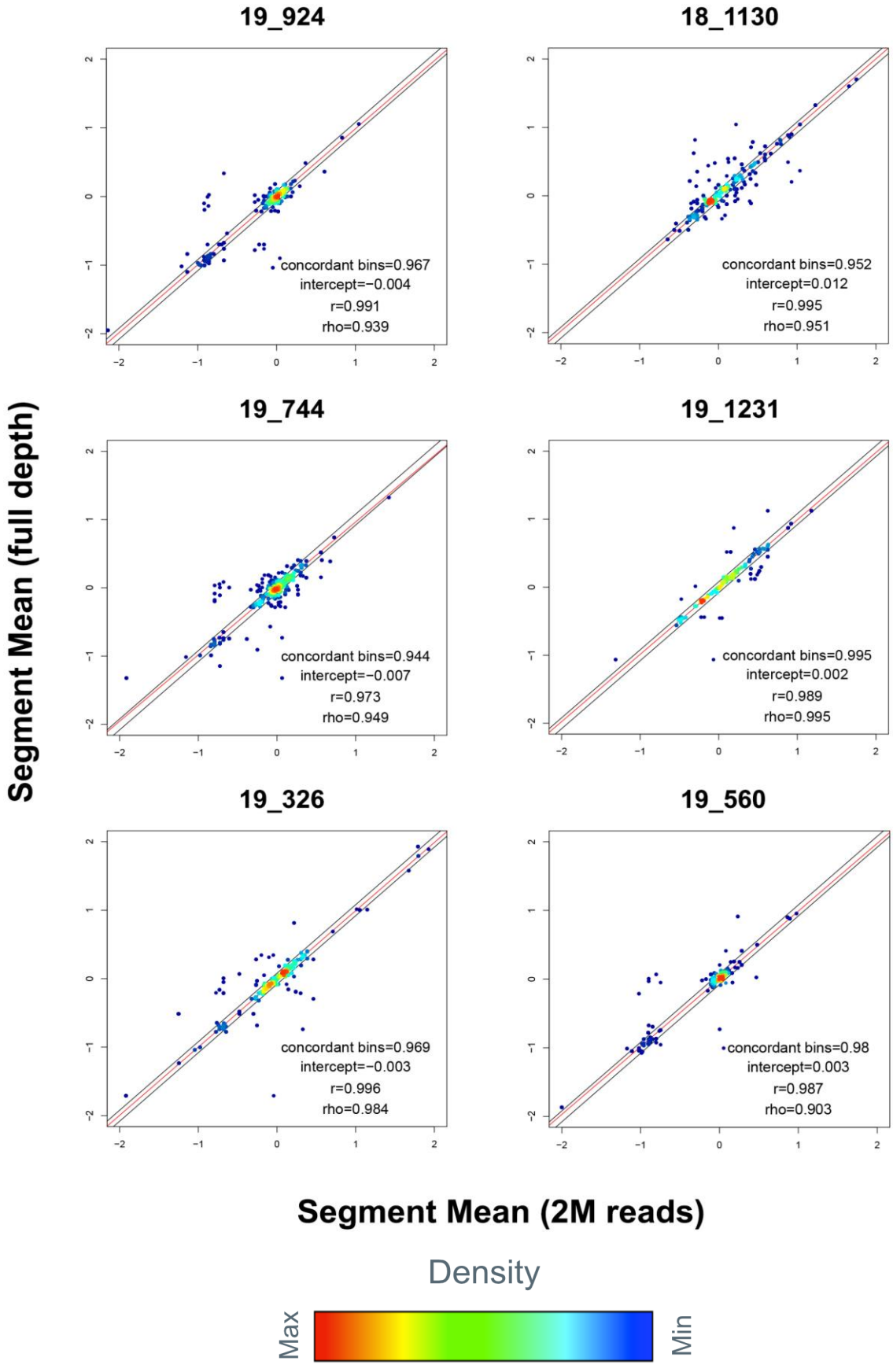

Figure 19: Comparison of segmentation results of 2 Million reads subsets: Comparison of Nanopore segment mean values from the full-depth BAM file and from a 2M reads subsampled dataset. Each genomic bin is represented as a dot, colours indicate dot density. Regression lines are shown in red. Black lines indicate the thresholds for concordant bins.

Illumina and Nanopore result comparison

We then compared the performance of Nanopore sequencing with a standard SGS approach by analysing four of the tumoral samples through Illumina sequencing (17-24M, 150bp single end reads, see methods). Illumina and Nanopore results (*"nocontrol"* mode) were strongly correlated (R = 0.93 – 0.99, $p << 0.001$), with concordant log2ratio values at 95-98% of the genomic bins (Figure 18, Table 6 **A**). To assess the performances of our approach at even lower sequencing depth, we subsampled the BAMs to 2M raw reads: the results obtained are highly concordant with the full-depth BAMs (R = 0.93 – 0.99, $p << 0.001$, 94-99% concordant bins, Figure 19**,** Table 6).

The marginal loss of performance observed is comparable to the one obtained when subsampling Illumina data (Table 4).

Detection of lung cancer-related CNVs

Since the ultimate aim of the analysis is to obtain information on the tumour, we next assessed the status of genes commonly altered in lung cancer, selected from 6 papers (Figure 20) (181-187). Notably, using read-count based methods such as NanoGLADIATOR, it is challenging to define the expected read-count for diploidy in presence of a high number of CNVs. Due to this, taking in account this limitation, we used a stricter log2ratio threshold (± 0.10) for the assessment of amplifications/deletions, to avoid false positives. Pathogenetic CNVs were readily observed, with EGFR amplification prominently present in all samples, and most of other genes altered in at least two samples. Many of these structural alterations directly affect progression of the cancer and therapeutic options. For example, RICTOR amplification identifies a subgroup of lung cancer and its presence has been linked to the response to mTOR inhibitors (185). Similarly, MYC amplification confers resistance to

pictilisib in models and PIK3CA amplification is associated with resistance to PI3K inhibition (260, 261) in mammary tumors.

A

Case	Contol	*Nocontrol* mode								*Paired* mode							
		Whole genome				Copy number altered regions				Whole genome				Copy number altered regions			
		Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)
19_744	HMP	96,10	-0,01	1,07	0,95	93,59	-0,02	1,07	0,97	94,52	0,01	0,90	0,93	90,67	0,00	0,90	0,93
18_1130	HF1	96,54	0,02	0,73	0,96	93,24	0,05	0,69	0,95	92,16	0,03	1,01	0,92	88,55	0,03	1,01	0,95
19_924	HMP	95,63	-0,01	0,99	0,93	89,26	-0,04	0,97	0,83	93,75	0,01	0,98	0,88	85,82	0,00	0,97	0,90
19_1231	HF1	98,42	0,00	0,79	0,99	98,05	-0,01	0,79	0,99	96,06	0,01	1,02	0,98	95,21	0,01	1,02	0,98

B

Case	Contol	*Nocontrol* mode								*Paired* mode							
		Whole genome				Copy number altered regions				Whole genome				Copy number altered regions			
		Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)	Concordant bins (%)	regression line (intercept)	regression line (r)	Correlation (rho)
19_744	HMP	94,40	-0,01	0,97	0,95	90,40	-0,01	0,98	0,96	94,63	0,01	1,00	0,93	91,25	0,01	0,99	0,94
18_1130	HF1	95,16	0,01	0,99	0,95	92,04	0,02	0,99	0,97	96,30	0,01	0,98	0,97	94,07	0,01	0,98	0,96
19_924	HMP	96,69	0,00	0,99	0,94	93,00	-0,01	0,99	0,89	97,45	0,00	0,92	0,93	90,07	0,00	0,94	0,83
19_1231	HF1	99,50	0,00	0,99	1,00	99,35	0,00	0,99	0,99	99,59	0,00	0,93	0,99	99,49	0,00	0,93	0,99
19_560	HMP	97,98	0,00	0,99	0,90	87,03	0,01	1,00	0,96	99,32	0,00	0,94	0,87	78,94	0,00	0,93	0,67
19_326	HMP	96,88	0,00	1,00	0,98	95,06	0,00	1,00	0,97	97,69	0,00	0,98	0,98	96,66	0,00	0,98	0,98
HM1	-	98,22	0,00	0,98	0,77	83,91	-0,01	0,97	0,93	-	-	-	-	-	-	-	-
HM2	-	98,71	-0,01	0,97	0,77	85,67	-0,03	0,95	0,92	-	-	-	-	-	-	-	-
HM3	-	96,96	0,00	0,98	0,72	77,92	0,00	0,98	0,79	-	-	-	-	-	-	-	-
HF1	-	98,50	0,00	1,02	0,78	79,36	-0,01	1,02	0,17	-	-	-	-	-	-	-	-

Table 6: Correlation of segmentation results. (A) Correlation of Illumina and Nanopore results, (B) Correlation of Nanopore results: subsampled BAMs (2M reads) Vs full BAMs ("nocontrol" and "paired" mode).

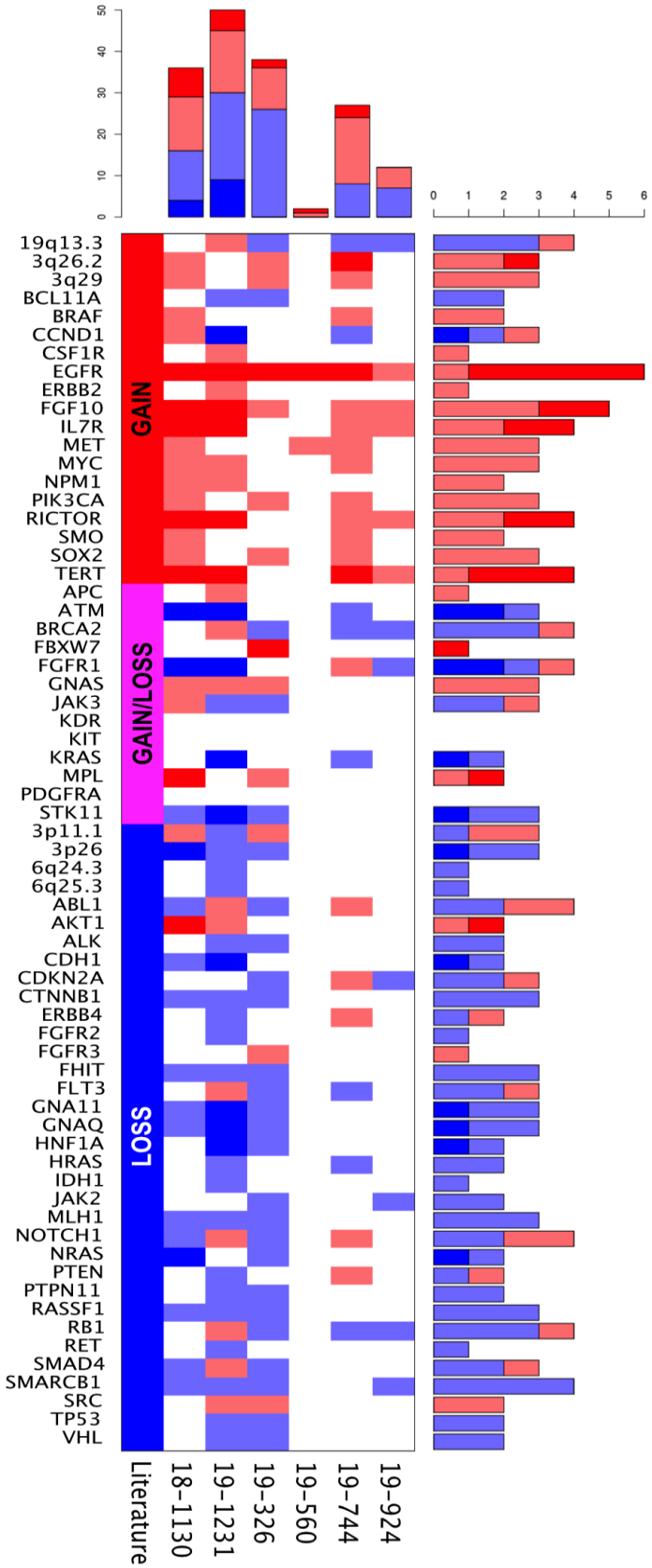

Figure 20 Landscape of clinically-relevant copy number variants. Copy number variants of specific genes (rows) are shown for the individual patients (columns). The shading indicates levels of amplification (red tones, 0.10-0.30, >0.30 log2ratio) and deletion (blue tones, 0.10-0.30, >0.30 negative log2ratio). The top and right bar plots show the number of CNVs in one patient and the number of patients with CNVs for a given gene, respectively. The expected status for a given gene based on the literature is shown in the left side.

Discussion

Our report is the first successful attempt to obtain a CNV profile from plasma cell-free DNA of cancer patients using Nanopore technology. Our results show that Nanopore sequencing has the same performance of SGS approaches and, in terms of throughput and sequencing costs, it is comparable to an Illumina MiSeq (V3 reagents, 22-25M single-end reads).

MinION is the entry-level sequencer by Nanopore technology and its cost is extremely low (~1000 euros) compared to SGS sequencers whose price ranges from tens to hundreds of thousands of euros. Reduced overall instrumentation costs makes this approach accessible to most of the research groups, which would otherwise be forced to outsource the sequencing, or to gain access to shared sequencers, leading often to long queues and delays. Moreover, SGS is cost effective only when dealing with a large number of patients. This aspect is crucial with regards to clinical analyses, as it leads to a centralization of sequencing-based assays, which are mainly performed in big hospitals that collect samples from larger geographic areas.

On the contrary, Nanopore technology is extremely scalable, and only a modest number of patients is required in a multiplexed run, leading to short recruitment times and, consequently, faster results.

As we demonstrate that reliable results can be obtained from as few as 2M reads. Based on the throughput obtained in our study, it should be possible to analyse up to 7-15 patients in a single run.

Since reads are stored as soon as they are produced, they can be analysed while the experiment is still running by taking advantage of the real-time mode of NanoGLADIATOR. This feature might come useful when analysing single samples, especially in those patients with lower fraction of ctDNA, for which a higher number or reads and, consequently, a higher resolution may be preferable. In such a context, it would be possible to inspect the CNV

profile while the run is still ongoing, and stop once the desired resolution is reached, saving the sequencing power of the flow cell, which can be washed and reused for other samples.

According to our sequencing statistics, 2M reads are produced in less than 3 hours. This means that the entire workflow -from blood withdrawal to bioinformatic analyses- can be performed in less than a working day. This is something unique to Nanopore sequencing, as SGS approaches based on sequence-by-synthesis technologies make reads available only at the end of the whole run, which can last days.

We have demonstrated that Nanopore sequencing for CNV analysis of short plasmatic cfDNA is feasible. Nanopore features represent advantages over current sequencing technologies, and might drive the adoption of molecular karyotyping from liquid biopsies as a tool for cancer monitoring in clinical settings. The applications of this approach are not limited to cancer and can be technically extended to other liquid biopsy-based fields such as noninvasive prenatal diagnosis. One limitation of this study is the lack of a comparison between histological samples and cfDNA, to assess if cfDNA is really representative of the tumor. On the other hand, this comparison is subject to a variety of biases such as: sampling biases of the tumoral tissue (the sampled portion may not completely reflect the entire CNV load of the tumor) and related to the fact that the patients may carry undetected metastasis, which contribute to the CNV profile observed in plasma samples, but would be ignored by tissue sampling. However, assessing the degree of reliability of cfDNA-based analyses is not the goal of the project.

In future, it will be interesting to assess the precision of the Nanopore approach by comparing multiple biological and technical replicates from the same patient.

Bibliography

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. The New England journal of medicine. 2020;382(13):1199-207.
2. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.
3. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nature reviews Microbiology. 2019;17(3):181-92.
4. Artika IM, Dewantari AK, Wiyatno A. Molecular biology of coronaviruses: current knowledge. Heliyon. 2020;6(8):e04743.
5. Masters PS. The molecular biology of coronaviruses. Advances in virus research. 2006;66:193-292.
6. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nature reviews Microbiology. 2016;14(8):523-34.
7. Chams N, Chams S, Badran R, Shams A, Araji A, Raad M, et al. COVID-19: A Multidisciplinary Review. Frontiers in public health. 2020;8:383.
8. Zheng J. SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat. International journal of biological sciences. 2020;16(10):1678-85.
9. Fan Y, Zhao K, Shi ZL, Zhou P. Bat Coronaviruses in China. Viruses. 2019;11(3).
10. Adachi S, Koma T, Doi N, Nomaguchi M, Adachi A. Commentary: Origin and evolution of pathogenic coronaviruses. Frontiers in immunology. 2020;11:811.
11. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature medicine. 2020;26(4):450-2.
12. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet (London, England). 2020;395(10224):565-74.
13. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of advanced research. 2020;24:91-8.
14. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3.
15. Huang Y, Yang C, Xu XF, Xu W, Liu SW. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta pharmacologica Sinica. 2020;41(9):1141-9.
16. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. Nature. 2020;581(7807):221-4.
17. Martínez-Hernández F, Isaak-Delgado AB, Alfonso-Toledo JA, Muñoz-García CI, Villalobos G, Aréchiga-Ceballos N, et al. Assessing the SARS-CoV-2 threat to wildlife: Potential risk to a broad range of mammals. Perspectives in ecology and conservation. 2020.
18. Mollica V, Rizzo A, Massari F. The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer. Future oncology (London, England). 2020;16(27):2029-33.
19. Glowacka I, Bertram S, Müller MA, Allen P, Soilleux E, Pfefferle S, et al. Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike

protein for membrane fusion and reduces viral control by the humoral immune response. Journal of virology. 2011;85(9):4122-34.
20. Sawicki SG, Sawicki DL, Younker D, Meyer Y, Thiel V, Stokes H, et al. Functional and genetic analysis of coronavirus replicase-transcriptase proteins. PLoS pathogens. 2005;1(4):e39.
21. Krichel B, Falke S, Hilgenfeld R, Redecke L, Uetrecht C. Processing of the SARS-CoV pp1a/ab nsp7-10 region. The Biochemical journal. 2020;477(5):1009-19.
22. Hsu MF, Kuo CJ, Chang KT, Chang HC, Chou CC, Ko TP, et al. Mechanism of the maturation process of SARS-CoV 3CL protease. The Journal of biological chemistry. 2005;280(35):31257-66.
23. Knoops K, Kikkert M, Worm SH, Zevenhoven-Dobbe JC, van der Meer Y, Koster AJ, et al. SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. PLoS biology. 2008;6(9):e226.
24. Jain J, Gaur S, Chaudhary Y, Kaul R. The molecular biology of intracellular events during Coronavirus infection cycle. Virusdisease. 2020;31(2):1-5.
25. Hagemeijer MC, Monastyrska I, Griffith J, van der Sluijs P, Voortman J, van Bergen en Henegouwen PM, et al. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. Virology. 2014;458-459:125-35.
26. Züst R, Cervantes-Barragán L, Kuri T, Blakqori G, Weber F, Ludewig B, et al. Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. PLoS pathogens. 2007;3(8):e109.
27. Wang Y, Grunewald M, Perlman S. Coronaviruses: An Updated Overview of Their Replication and Pathogenesis. Methods in molecular biology (Clifton, NJ). 2020;2203:1-29.
28. Djikeng A, Spiro D. Advancing full length genome sequencing for human RNA viral pathogens. Future virology. 2009;4(1):47-53.
29. Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. Journal of medical virology. 2020;92(4):433-40.
30. Rowe CL, Fleming JO, Nathan MJ, Sgro JY, Palmenberg AC, Baker SC. Generation of coronavirus spike deletion variants by high-frequency recombination at regions of predicted RNA secondary structure. Journal of virology. 1997;71(8):6183-90.
31. Rosenberg R, Johansson MA, Powers AM, Miller BR. Search strategy has influenced the discovery rate of human viruses. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(34):13961-4.
32. Rosenberg R. Detecting the emergence of novel, zoonotic viruses pathogenic to humans. Cellular and molecular life sciences : CMLS. 2015;72(6):1115-25.
33. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA biology. 2011;8(2):270-9.
34. Velavan TP, Meyer CG. The COVID-19 epidemic. Tropical medicine & international health : TM & IH. 2020;25(3):278-80.
35. Perlman S. Another Decade, Another Coronavirus. The New England journal of medicine. 2020;382(8):760-2.
36. Lam TT, Jia N, Zhang YW, Shum MH, Jiang JF, Zhu HC, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature. 2020;583(7815):282-5.
37. Rahman S, Hoque N, Islam R, Akter S, Rubayet-Ul-Alam ASM, Siddique MA, et al. Epitope-based chimeric peptide vaccine design against S, M and E proteins of SARS-CoV-2 etiologic agent of global pandemic COVID-19: an in silico approach. bioRxiv; 2020.

38. Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science (New York, NY). 2020;368(6491):630-3.
39. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell discovery. 2020;6:14.
40. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review. 2020;7(6):1012-23.
41. Ren LL, Wang YM, Wu ZQ, Xiang ZC, Guo L, Xu T, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chinese medical journal. 2020;133(9):1015-24.
42. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(17):9241-3.
43. Mossel EC, Wang J, Jeffers S, Edeen KE, Wang S, Cosgrove GP, et al. SARS-CoV replicates in primary human alveolar type II cell cultures but not in type I-like cells. Virology. 2008;372(1):127-35.
44. Mason RJ. Pathogenesis of COVID-19 from a cell biology perspective. The European respiratory journal. 2020;55(4).
45. Esmaeilzadeh A, Elahi R. Immunobiology and immunotherapy of COVID-19: A clinically updated overview. Journal of cellular physiology. 2020.
46. Jia X, Yin C, Lu S, Chen Y, Liu Q, Bai J, et al. Two Things about COVID-19 Might Need Attention. Preprints.org; 2020.
47. Kuba K, Imai Y, Penninger JM. Angiotensin-converting enzyme 2 in lung diseases. Current opinion in pharmacology. 2006;6(3):271-6.
48. Imai Y, Kuba K, Rao S, Huan Y, Guo F, Guan B, et al. Angiotensin-converting enzyme 2 protects from severe acute lung failure. Nature. 2005;436(7047):112-6.
49. Li X, Geng M, Peng Y, Meng L, Lu S. Molecular immune pathogenesis and diagnosis of COVID-19. Journal of pharmaceutical analysis. 2020;10(2):102-8.
50. Li G, Fan Y, Lai Y, Han T, Li Z, Zhou P, et al. Coronavirus infections and immune responses. Journal of medical virology. 2020;92(4):424-32.
51. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England). 2020;395(10223):497-506.
52. Vabret N, Britton GJ, Gruber C, Hegde S, Kim J, Kuksin M, et al. Immunology of COVID-19: Current State of the Science. Immunity. 2020;52(6):910-41.
53. Chiappelli F, Khakshooy A, Greenberg G. CoViD-19 Immunopathology and Immunotherapy. Bioinformation. 2020;16(3):219-22.
54. Zhang F, Gan R, Zhen Z, Hu X, Li X, Zhou F, et al. Adaptive immune responses to SARS-CoV-2 infection in severe versus mild individuals. Signal transduction and targeted therapy. 2020;5(1):156.
55. Chen SH, Habib G, Yang CY, Gu ZW, Lee BR, Weng SA, et al. Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. Science (New York, NY). 1987;238(4825):363-6.
56. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell. 1987;50(6):831-40.
57. Shaw JM, Feagin JE, Stuart K, Simpson L. Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell. 1988;53(3):401-11.

58. Benne R. RNA editing in trypanosomes. European journal of biochemistry. 1994;221(1):9-23.
59. Mahendran R, Spottswood MS, Ghate A, Ling ML, Jeng K, Miller DL. Editing of the mitochondrial small subunit rRNA in Physarum polycephalum. The EMBO journal. 1994;13(1):232-40.
60. Grosjean H, Auxilien S, Constantinesco F, Simon C, Corda Y, Becker HF, et al. Enzymatic conversion of adenosine to inosine and to N1-methylinosine in transfer RNAs: a review. Biochimie. 1996;78(6):488-501.
61. Yoshizawa S, Fourmy D, Puglisi JD. Recognition of the codon-anticodon helix by ribosomal RNA. Science (New York, NY). 1999;285(5434):1722-5.
62. Teng B, Burant Cf Fau - Davidson NO, Davidson NO. Molecular cloning of an apolipoprotein B messenger RNA editing protein. (0036-8075 (Print)).
63. Hirano K, Young SG, Farese RV, Jr., Ng J, Sande E, Warburton C, et al. Targeted disruption of the mouse apobec-1 gene abolishes apolipoprotein B mRNA editing and eliminates apolipoprotein B48. The Journal of biological chemistry. 1996;271(17):9887-90.
64. Yamanaka S, Poksay KS, Arnold KS, Innerarity TL. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. Genes & development. 1997;11(3):321-33.
65. Blanc V, Park E, Schaefer S, Miller M, Lin Y, Kennedy S, et al. Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. Genome biology. 2014;15(6):R79.
66. Rosenberg BR, Hamilton CE, Mwangi MM, Dewell S, Papavasiliou FN. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nature structural & molecular biology. 2011;18(2):230-6.
67. Harjanto D, Papamarkou T, Oates CJ, Rayon-Estrada V, Papavasiliou FN, Papavasiliou A. RNA editing generates cellular subsets with diverse sequence within populations. Nature communications. 2016;7:12145.
68. Rayon-Estrada V, Harjanto D, Hamilton CE, Berchiche YA, Gantman EC, Sakmar TP, et al. Epitranscriptomic profiling across cell types reveals associations between APOBEC1-mediated RNA editing, gene expression outcomes, and cellular function. Proceedings of the National Academy of Sciences of the United States of America. 2017;114(50):13296-301.
69. Conticello SG, Thomas CJ, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Molecular biology and evolution. 2005;22(2):367-77.
70. Sharma S, Patnaik SK, Kemer Z, Baysal BE. Transient overexpression of exogenous APOBEC3A causes C-to-U RNA editing of thousands of genes. RNA biology. 2017;14(5):603-10.
71. Sharma S, Patnaik SK, Taggart RT, Kannisto ED, Enriquez SM, Gollnick P, et al. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nature communications. 2015;6:6881.
72. Sharma S, Wang J, Alqassim E, Portwood S, Cortes Gomez E, Maguire O, et al. Mitochondrial hypoxic stress induces widespread RNA editing by APOBEC3G in natural killer cells. Genome biology. 2019;20(1):37.
73. Vartanian JP, Meyerhans A, Asjö B, Wain-Hobson S. Selection, recombination, and G----A hypermutation of human immunodeficiency virus type 1 genomes. Journal of virology. 1991;65(4):1779-88.

74. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, Watt IN, et al. DNA deamination mediates innate immunity to retroviral infection. Cell. 2003;113(6):803-9.
75. Mahieux R, Suspène R, Delebecque F, Henry M, Schwartz O, Wain-Hobson S, et al. Extensive editing of a small fraction of human T-cell leukemia virus type 1 genomes by four APOBEC3 cytidine deaminases. The Journal of general virology. 2005;86(Pt 9):2489-94.
76. Noguchi C, Ishino H, Tsuge M, Fujimoto Y, Imamura M, Takahashi S, et al. G to A hypermutation of hepatitis B virus. Hepatology (Baltimore, Md). 2005;41(3):626-33.
77. Vartanian JP, Guétard D, Henry M, Wain-Hobson S. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. Science (New York, NY). 2008;320(5873):230-3.
78. Suspène R, Aynaud MM, Koch S, Pasdeloup D, Labetoulle M, Gaertner B, et al. Genetic editing of herpes simplex virus 1 and Epstein-Barr herpesvirus genomes by human APOBEC3 cytidine deaminases in culture and in vivo. Journal of virology. 2011;85(15):7594-602.
79. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely High Mutation Rate of HIV-1 In Vivo. PLoS biology. 2015;13(9):e1002251.
80. Peretti A, Geoghegan EM, Pastrana DV, Smola S, Feld P, Sauter M, et al. Characterization of BK Polyomaviruses from Kidney Transplant Recipients Suggests a Role for APOBEC3 in Driving In-Host Virus Evolution. Cell host & microbe. 2018;23(5):628-35.e7.
81. Perelygina L, Chen MH, Suppiah S, Adebayo A, Abernathy E, Dorsey M, et al. Infectious vaccine-derived rubella viruses emerge, persist, and evolve in cutaneous granulomas of children with primary immunodeficiencies. PLoS pathogens. 2019;15(10):e1008080.
82. Bass BL, Weintraub H. An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell. 1988;55(6):1089-98.
83. Kim U, Garner TL, Sanford T, Speicher D, Murray JM, Nishikura K. Purification and characterization of double-stranded RNA adenosine deaminase from bovine nuclear extracts. The Journal of biological chemistry. 1994;269(18):13480-9.
84. Kim U, Wang Y, Sanford T, Zeng Y, Nishikura K. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(24):11457-61.
85. Melcher T, Maas S, Herb A, Sprengel R, Seeburg PH, Higuchi M. A mammalian RNA editing enzyme. Nature. 1996;379(6564):460-4.
86. Keegan LP, McGurk L, Palavicini JP, Brindle J, Paro S, Li X, et al. Functional conservation in human and Drosophila of Metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects. Nucleic acids research. 2011;39(16):7249-62.
87. Paul MS, Bass BL. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. The EMBO journal. 1998;17(4):1120-7.
88. Behm M, Öhman M. RNA Editing: A Contributor to Neuronal Dynamics in the Mammalian Brain. Trends in genetics : TIG. 2016;32(3):165-75.
89. Yablonovitch AL, Deng P, Jacobson D, Li JB. The evolution and adaptation of A-to-I RNA editing. PLoS genetics. 2017;13(11):e1007064.
90. Neeman Y, Levanon EY, Jantsch MF, Eisenberg E. RNA editing level in the mouse is determined by the genomic repeat repertoire. RNA (New York, NY). 2006;12(10):1802-9.

91. Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, Pesole G. Profiling RNA editing in human tissues: towards the inosinome Atlas. Scientific reports. 2015;5:14941.
92. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. Nature methods. 2012;9(6):579-81.
93. Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science (New York, NY). 2007;315(5815):1137-40.
94. Ota H, Sakurai M, Gupta R, Valente L, Wulff BE, Ariyoshi K, et al. ADAR1 forms a complex with Dicer to promote microRNA processing and RNA-induced gene silencing. Cell. 2013;153(3):575-89.
95. Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, et al. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nature structural & molecular biology. 2006;13(1):13-21.
96. Kawahara Y, Zinshteyn B, Chendrimada TP, Shiekhattar R, Nishikura K. RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer-TRBP complex. EMBO reports. 2007;8(8):763-9.
97. Bazak L, Haviv A, Barak M, Jacob-Hirsch J, Deng P, Zhang R, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome research. 2014;24(3):365-76.
98. Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic acids research. 2017;45(D1):D750-d7.
99. Rosenthal JJ. The emerging role of RNA editing in plasticity. The Journal of experimental biology. 2015;218(Pt 12):1812-21.
100. Levanon EY, Eisenberg E. Does RNA editing compensate for Alu invasion of the primate genome? BioEssays : news and reviews in molecular, cellular and developmental biology. 2015;37(2):175-81.
101. DeCerbo J, Carmichael GG. Retention and repression: fates of hyperedited RNAs in the nucleus. Current opinion in cell biology. 2005;17(3):302-8.
102. Mehedi M, Hoenen T, Robertson S, Ricklefs S, Dolan MA, Taylor T, et al. Ebola virus RNA editing depends on the primary editing site sequence and an upstream secondary structure. PLoS pathogens. 2013;9(10):e1003677.
103. Tomaselli S, Galeano F, Locatelli F, Gallo A. ADARs and the Balance Game between Virus Infection and Innate Immune Cell Response. Current issues in molecular biology. 2015;17:37-51.
104. Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, et al. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science (New York, NY). 2015;349(6252):1115-20.
105. Mannion NM, Greenwood SM, Young R, Cox S, Brindle J, Read D, et al. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. Cell reports. 2014;9(4):1482-94.
106. Pestal K, Funk CC, Snyder JM, Price ND, Treuting PM, Stetson DB. Isoforms of RNA-Editing Enzyme ADAR1 Independently Control Nucleic Acid Sensor MDA5-Driven Autoimmunity and Multi-organ Development. Immunity. 2015;43(5):933-44.
107. George CX, Gan Z, Liu Y, Samuel CE. Adenosine deaminases acting on RNA, RNA editing, and interferon action. Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research. 2011;31(1):99-117.

108. Sarvestani ST, Tate MD, Moffat JM, Jacobi AM, Behlke MA, Miller AR, et al. Inosine-mediated modulation of RNA sensing by Toll-like receptor 7 (TLR7) and TLR8. Journal of virology. 2014;88(2):799-810.
109. Yang S, Deng P, Zhu Z, Zhu J, Wang G, Zhang L, et al. Adenosine deaminase acting on RNA 1 limits RIG-I RNA detection and suppresses IFN production responding to viral and endogenous RNAs. Journal of immunology (Baltimore, Md : 1950). 2014;193(7):3436-45.
110. Moris A, Murray S, Cardinaud S. AID and APOBECs span the gap between innate and adaptive immunity. Frontiers in microbiology. 2014;5:534.
111. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific reports. 2018;8(1):10950.
112. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-60.
113. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-9.
114. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome research. 2012;22(3):568-76.
115. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013;31(3):213-9.
116. Picardi E, Pesole G. REDItools: high-throughput RNA editing detection made easy. Bioinformatics (Oxford, England). 2013;29(14):1813-4.
117. Piechotta M, Wyler E, Ohler U, Landthaler M, Dieterich C. JACUSA: site-specific identification of RNA editing events from replicate sequencing data. BMC bioinformatics. 2017;18(1):7.
118. Xu C, Nezami Ranjbar MR, Wu Z, DiCarlo J, Wang Y. Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller. BMC genomics. 2017;18(1):5.
119. Dua K, Shukla SD, Hansbro PM. Aspiration techniques for bronchoalveolar lavage in translational respiratory research: Paving the way to develop novel therapeutic moieties. Journal of biological methods. 2017;4(3):e73.
120. Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, et al. Detection of SARS-CoV-2 in Different Types of Clinical Specimens. Jama. 2020;323(18):1843-4.
121. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). 2014;30(15):2114-20.
122. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics (Oxford, England). 2016;32(2):292-4.
123. Flati T, Gioiosa S, Spallanzani N, Tagliaferri I, Diroma MA, Pesole G, et al. HPC-REDItools: a novel HPC-aware tool for improved large scale RNA-editing analysis. BMC bioinformatics. 2020;21(Suppl 10):353.
124. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS computational biology. 2013;9(8):e1003118.
125. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004;32(5):1792-7.

126. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerging microbes & infections. 2020;9(1):313-9.
127. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome-Coronavirus 2 in Patients With Coronavirus Disease 2019. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2020;71(15):713-20.
128. Lo Giudice C, Tangaro MA, Pesole G, Picardi E. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. Nature protocols. 2020;15(3):1098-131.
129. Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, et al. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS pathogens. 2010;6(5):e1000896.
130. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS pathogens. 2013;9(8):e1003565.
131. Rosani U, Bai CM, Maso L, Shapiro M, Abbadi M, Domeneghetti S, et al. A-to-I editing of Malacoherpesviridae RNAs supports the antiviral role of ADAR1 in mollusks. BMC evolutionary biology. 2019;19(1):149.
132. Carpenter JA, Keegan LP, Wilfert L, O'Connell MA, Jiggins FM. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). BMC genetics. 2009;10:75.
133. Zahn RC, Schelp I, Utermöhlen O, von Laer D. A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus. Journal of virology. 2007;81(2):457-64.
134. Bar-Yaacov D, Avital G, Levin L, Richards AL, Hachen N, Rebolledo Jaramillo B, et al. RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA. Genome research. 2013;23(11):1789-96.
135. Lehmann KA, Bass BL. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry. 2000;39(42):12875-84.
136. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science (New York, NY). 2009;324(5931):1210-3.
137. Bahn JH, Lee JH, Li G, Greer C, Peng G, Xiao X. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome research. 2012;22(1):142-50.
138. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nature communications. 2011;2(1):319.
139. Lerner T, Papavasiliou FN, Pecori R. RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-to-U RNA Editing Machinery and Its Implication in Human Disease. Genes. 2018;10(1).
140. Roth SH, Levanon EY, Eisenberg E. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. Nature methods. 2019;16(11):1131-8.
141. Wilson BD, Eisenstein M, Soh HT. High-Fidelity Nanopore Sequencing of Ultra-Short DNA Targets. Analytical chemistry. 2019;91(10):6783-9.
142. Osenberg S, Dominissini D, Rechavi G, Eisenberg E. Widespread cleavage of A-to-I hyperediting substrates. RNA (New York, NY). 2009;15(9):1632-9.
143. Ko NL, Birlouez E, Wain-Hobson S, Mahieux R, Vartanian JP. Hyperediting of human T-cell leukemia virus type 2 and simian T-cell leukemia virus type 3 by the dsRNA adenosine deaminase ADAR-1. The Journal of general virology. 2012;93(Pt 12):2646-51.

144. Porath HT, Carmi S, Levanon EY. A genome-wide map of hyper-edited RNA reveals numerous new sites. Nature communications. 2014;5:4726.
145. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181(4):914-21.e10.
146. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. Population stratification of a common APOBEC gene deletion polymorphism. PLoS genetics. 2007;3(4):e63.
147. Long J, Delahanty RJ, Li G, Gao YT, Lu W, Cai Q, et al. A common deletion in the APOBEC3 genes and breast cancer risk. Journal of the National Cancer Institute. 2013;105(8):573-9.
148. Fanciulli M, Petretto E, Aitman TJ. Gene copy number variation and common human disease. Clinical genetics. 2010;77(3):201-13.
149. Shlien A, Malkin D. Copy number variations and cancer. Genome medicine. 2009;1(6):62.
150. Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, et al. Segmental copy number variation shapes tissue transcriptomes. Nature genetics. 2009;41(4):424-9.
151. Cahan P, Li Y, Izumi M, Graubert TA. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. Nature genetics. 2009;41(4):430-7.
152. Holt R, Sykes NH, Conceição IC, Cazier J-B, Anney RJL, Oliveira G, et al. CNVs leading to fusion transcripts in individuals with autism spectrum disorder. European Journal of Human Genetics. 2012;20(11):1141-7.
153. Bittel DC, Butler MG. Prader-Willi syndrome: clinical genetics, cytogenetics and molecular biology. Expert reviews in molecular medicine. 2005;7(14):1-20.
154. Meyer-Lindenberg A, Mervis CB, Berman KF. Neural mechanisms in Williams syndrome: a unique window to genetic influences on cognition and behaviour. Nature reviews Neuroscience. 2006;7(5):380-93.
155. Elsea SH, Girirajan S. Smith-Magenis syndrome. European journal of human genetics : EJHG. 2008;16(4):412-21.
156. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science (New York, NY). 2005;307(5714):1434-40.
157. Aitman TJ, Dong R, Vyse TJ, Norsworthy PJ, Johnson MD, Smith J, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006;439(7078):851-5.
158. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nature genetics. 2007;39(6):721-3.
159. Willcocks LC, Lyons PA, Clatworthy MR, Robinson JI, Yang W, Newland SA, et al. Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. The Journal of experimental medicine. 2008;205(7):1573-82.
160. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. American journal of human genetics. 2007;80(6):1037-54.
161. McKinney C, Merriman ME, Chapman PT, Gow PJ, Harrison AA, Highton J, et al. Evidence for an influence of chemokine ligand 3-like 1 (CCL3L1) gene copy number on

susceptibility to rheumatoid arthritis. Annals of the rheumatic diseases. 2008;67(3):409-13.
162. Burns JC, Shimizu C, Gonzalez E, Kulkarni H, Patel S, Shike H, et al. Genetic variations in the receptor-ligand pair CCR5 and CCL3L1 are important determinants of susceptibility to Kawasaki disease. The Journal of infectious diseases. 2005;192(2):344-9.
163. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, et al. A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. American journal of human genetics. 2006;79(3):439-48.
164. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nature genetics. 2008;40(9):1107-12.
165. Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nature genetics. 2008;40(1):23-5.
166. Piirilä P, Wikman H, Luukkonen R, Kääriä K, Rosenberg C, Nordman H, et al. Glutathione S-transferase genotypes and allergic responses to diisocyanate exposure. Pharmacogenetics. 2001;11(5):437-45.
167. Ivaschenko TE, Sideleva OG, Baranov VS. Glutathione- S-transferase micro and theta gene polymorphisms as new risk factors of atopic bronchial asthma. Journal of molecular medicine (Berlin, Germany). 2002;80(1):39-43.
168. Brasch-Andersen C, Christiansen L, Tan Q, Haagerup A, Vestbo J, Kruse TA. Possible gene dosage effect of glutathione-S-transferases on atopic asthma: using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. Human mutation. 2004;24(3):208-14.
169. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science (New York, NY). 2007;316(5823):445-9.
170. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ, et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nature genetics. 2007;39(3):319-28.
171. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, et al. Association between microdeletion and microduplication at 16p11.2 and autism. The New England journal of medicine. 2008;358(7):667-75.
172. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. Nature genetics. 2008;40(7):880-5.
173. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science (New York, NY). 2008;320(5875):539-43.
174. Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;455(7210):232-6.
175. Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, et al. alpha-Synuclein locus triplication causes Parkinson's disease. Science (New York, NY). 2003;302(5646):841.

176. Ibáñez P, Bonnet AM, Débarges B, Lohmann E, Tison F, Pollak P, et al. Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. Lancet (London, England). 2004;364(9440):1169-71.
177. Veldink JH, van den Berg LH, Cobben JM, Stulp RP, De Jong JM, Vogels OJ, et al. Homozygous deletion of the survival motor neuron 2 gene is a prognostic factor in sporadic ALS. Neurology. 2001;56(6):749-52.
178. Wang J, Ban MR, Hegele RA. Multiplex ligation-dependent probe amplification of LDLR enhances molecular diagnosis of familial hypercholesterolemia. Journal of lipid research. 2005;46(2):366-72.
179. Lanktree M, Hegele RA. Copy number variation in metabolic phenotypes. Cytogenetic and genome research. 2008;123(1-4):169-75.
180. Tosi I, Toledo-Leiva P, Neuwirth C, Naoumova RP, Soutar AK. Genetic defects causing familial hypercholesterolaemia: identification of deletions and duplications in the LDL-receptor gene and summary of all mutations found in patients attending the Hammersmith Hospital Lipid Clinic. Atherosclerosis. 2007;194(1):102-11.
181. Liao Y, Ma Z, Zhang Y, Li D, Lv D, Chen Z, et al. Targeted deep sequencing from multiple sources demonstrates increased NOTCH1 alterations in lung cancer patient plasma. Cancer medicine. 2019;8(12):5673-86.
182. Peng H, Lu L, Zhou Z, Liu J, Zhang D, Nan K, et al. CNV Detection from Circulating Tumor DNA in Late Stage Non-Small Cell Lung Cancer Patients. Genes. 2019;10(11).
183. Chen X, Chang CW, Spoerke JM, Yoh KE, Kapoor V, Baudo C, et al. Low-pass Whole-genome Sequencing of Circulating Cell-free DNA Demonstrates Dynamic Changes in Genomic Copy Number in a Squamous Lung Cancer Clinical Cohort. Clinical cancer research : an official journal of the American Association for Cancer Research. 2019;25(7):2254-63.
184. Vanhecke E, Valent A, Tang X, Vielh P, Friboulet L, Tang T, et al. 19q13-ERCC1 gene copy number increase in non--small-cell lung cancer. Clinical lung cancer. 2013;14(5):549-57.
185. Sakre N, Wildey G, Behtaj M, Kresak A, Yang M, Fu P, et al. RICTOR amplification identifies a subgroup in small cell lung cancer and predicts response to drugs targeting mTOR. Oncotarget. 2017;8(4):5992-6002.
186. Bowcock AM. DNA copy number changes as diagnostic tools for lung cancer. Thorax. 2014;69(5):496.
187. Du M, Thompson J, Fisher H, Zhang P, Huang CC, Wang L. Genomic alterations of plasma cell-free DNAs in small cell lung cancer and their clinical relevance. Lung cancer (Amsterdam, Netherlands). 2018;120:113-21.
188. Leary RJ, Lin JC, Cummins J, Boca S, Wood LD, Parsons DW, et al. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(42):16224-9.
189. Paris PL, Andaya A, Fridlyand J, Jain AN, Weinberg V, Kowbel D, et al. Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. Human molecular genetics. 2004;13(13):1303-13.
190. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.
191. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nature Reviews Cancer. 2020;20(10):555-72.

192. Knudson AG, Jr. Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences of the United States of America. 1971;68(4):820-3.
193. Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, et al. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(32):11264-9.
194. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature reviews Genetics. 2009;10(8):551-64.
195. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. Trends in cancer. 2016;2(1):49-63.
196. El-Deiry WS, Taylor B, Neal JW. Tumor Evolution, Heterogeneity, and Therapy for Our Patients With Advanced Cancer: How Far Have We Come? American Society of Clinical Oncology educational book American Society of Clinical Oncology Annual Meeting. 2017;37:e8-e15.
197. Marine JC, Dawson SJ, Dawson MA. Non-genetic mechanisms of therapeutic resistance in cancer. Nature reviews Cancer. 2020.
198. Zevallos A, Bravo L, Bretel D, Paez K, Infante U, Cárdenas N, et al. The hispanic landscape of triple negative breast cancer. Critical reviews in oncology/hematology. 2020;155:103094.
199. Wu SG, Chiang CL, Liu CY, Wang CC, Su PL, Hsia TC, et al. An Observational Study of Acquired EGFR T790M-Dependent Resistance to EGFR-TKI Treatment in Lung Adenocarcinoma Patients in Taiwan. Frontiers in oncology. 2020;10:1481.
200. Jayaram A, Wingate A, Wetterskog D, Conteduca V, Khalaf D, Sharabiani MTA, et al. Plasma Androgen Receptor Copy Number Status at Emergence of Metastatic Castration-Resistant Prostate Cancer: A Pooled Multicohort Analysis. JCO precision oncology. 2019;3.
201. Yildiz OG, Aslan D, Akalin H, Erdem Y, Canoz O, Aytekin A, et al. The Effects of O(6)-methyl Guanine DNA-methyl Transferase Promotor Methylation and CpG1, CpG2, CpG3 and CpG4 Methylation on Treatment Response and their Prognostic Significance in Patients with Glioblastoma. Balkan journal of medical genetics : BJMG. 2020;23(1):33-41.
202. Aguilar-Mahecha A, Lafleur J, Pelmus M, Seguin C, Lan C, Discepola F, et al. The identification of challenges in tissue collection for biomarker studies: the Q-CROC-03 neoadjuvant breast cancer translational trial experience. Modern Pathology. 2017;30(11):1567-76.
203. Cree IA. Liquid biopsy for cancer patients: Principles and practice. Pathogenesis. 2015;2(1):1-4.
204. Kunju LP, Carskadon S, Siddiqui J, Tomlins SA, Chinnaiyan AM, Palanisamy N. Novel RNA hybridization method for the in situ detection of ETV1, ETV4, and ETV5 gene fusions in prostate cancer. Applied immunohistochemistry & molecular morphology : AIMM. 2014;22(8):e32-40.
205. Warrick JI, Tomlins SA, Carskadon SL, Young AM, Siddiqui J, Wei JT, et al. Evaluation of tissue PCA3 expression in prostate cancer by RNA in situ hybridization--a correlative study with urine PCA3 and TMPRSS2-ERG. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc. 2014;27(4):609-20.

206. Carvajal-Hausdorf DE, Schalper KA, Neumeister VM, Rimm DL. Quantitative measurement of cancer tissue biomarkers in the lab and in the clinic. Laboratory investigation; a journal of technical methods and pathology. 2015;95(4):385-96.
207. Gupta S, Vanderbilt C, Abida W, Fine SW, Tickoo SK, Al-Ahmadie HA, et al. Immunohistochemistry-based assessment of androgen receptor status and the AR-null phenotype in metastatic castrate resistant prostate cancer. Prostate cancer and prostatic diseases. 2020;23(3):507-16.
208. Bozzetti C, Nizzoli R, Tiseo M, Squadrilli A, Lagrasta C, Buti S, et al. ALK and ROS1 rearrangements tested by fluorescence in situ hybridization in cytological smears from advanced non-small cell lung cancer patients. Diagnostic cytopathology. 2015;43(11):941-6.
209. Vuong HG, Nguyen TQ, Ngo TNM, Nguyen HC, Fung KM, Dunn IF. The interaction between TERT promoter mutation and MGMT promoter methylation on overall survival of glioma patients: a meta-analysis. BMC cancer. 2020;20(1):897.
210. Plaska SW, Liu CJ, Lim JS, Rege J, Bick NR, Lerario AM, et al. Targeted RNAseq of Formalin-Fixed Paraffin-Embedded Tissue to Differentiate Among Benign and Malignant Adrenal Cortical Tumors. Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme. 2020;52(8):607-13.
211. María A, Matías AÁ, Carmen B. Liquid biopsy for cancer management: a revolutionary but still limited new tool for precision medicine. Advances in Laboratory Medicine / Avances en Medicina de Laboratorio. 2020;1(3):20200009.
212. Pettit SJ, Seymour K, O'Flaherty E, Kirby JA. Immune selection in neoplasia: towards a microevolutionary model of cancer development. British journal of cancer. 2000;82(12):1900-6.
213. Szilágyi M, Pös O, Márton É, Buglyó G, Soltész B, Keserű J, et al. Circulating Cell-Free Nucleic Acids: Main Characteristics and Clinical Application. International journal of molecular sciences. 2020;21(18).
214. Cervena K, Vodicka P, Vymetalkova V. Diagnostic and prognostic impact of cell-free DNA in human cancers: Systematic review. Mutation research. 2019;781:100-29.
215. Fang R, Zhu Y, Khadka VS, Zhang F, Jiang B, Deng Y. The Evaluation of Serum Biomarkers for Non-small Cell Lung Cancer (NSCLC) Diagnosis. Frontiers in physiology. 2018;9:1710.
216. Pron G. Prostate-Specific Antigen (PSA)-Based Population Screening for Prostate Cancer: An Evidence-Based Analysis. Ontario health technology assessment series. 2015;15(10):1-64.
217. Howard J, Wyse C, Argyle D, Quinn C, Kelly P, McCann A. Exosomes as Biomarkers of Human and Feline Mammary Tumours; A Comparative Medicine Approach to Unravelling the Aggressiveness of TNBC. Biochimica et biophysica acta Reviews on cancer. 2020;1874(2):188431.
218. Goodman C, Speers CW. The role of circulating tumor cells in breast cancer and implications for radiation treatment decisions. International journal of radiation oncology, biology, physics. 2020.
219. Martignano F. Cell-Free DNA: An Overview of Sample Types and Isolation Procedures. Methods in molecular biology (Clifton, NJ). 2019;1909:13-27.
220. Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(40):E5503-12.

221. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016;164(1-2):57-68.
222. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. Cancer research. 1977;37(3):646-50.
223. Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017;545(7655):446-51.
224. Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology. 1989;46(5):318-22.
225. De Mattos-Arruda L, Weigelt B, Cortes J, Won HH, Ng CKY, Nuciforo P, et al. Capturing intra-tumor genetic heterogeneity by de novo mutation profiling of circulating cell-free tumor DNA: a proof-of-principle. Annals of oncology : official journal of the European Society for Medical Oncology. 2014;25(9):1729-35.
226. Jamal-Hanjani M, Wilson GA, Horswell S, Mitter R, Sakarya O, Constantin T, et al. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. Annals of oncology : official journal of the European Society for Medical Oncology. 2016;27(5):862-7.
227. Murtaza M, Dawson S-J, Pogrebniak K, Rueda OM, Provenzano E, Grant J, et al. Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. Nature communications. 2015;6(1):8760.
228. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Science translational medicine. 2014;6(224):224ra24.
229. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, et al. Circulating mutant DNA to assess tumor dynamics. Nature medicine. 2008;14(9):985-90.
230. El Messaoudi S, Mouliere F, Du Manoir S, Bascoul-Mollevi C, Gillet B, Nouaille M, et al. Circulating DNA as a Strong Multimarker Prognostic Tool for Metastatic Colorectal Cancer Patient Management Care. Clinical cancer research : an official journal of the American Association for Cancer Research. 2016;22(12):3067-77.
231. Thierry AR, Mouliere F, El Messaoudi S, Mollevi C, Lopez-Crapez E, Rolet F, et al. Clinical validation of the detection of KRAS and BRAF mutations from circulating tumor DNA. Nature medicine. 2014;20(4):430-5.
232. Ulz P, Auer M, Heitzer E. Detection of Circulating Tumor DNA in the Blood of Cancer Patients: An Important Tool in Cancer Chemoprevention. Methods in molecular biology (Clifton, NJ). 2016;1379:45-68.
233. Alimirzaie S, Bagherzadeh M, Akbari MR. Liquid biopsy in breast cancer: A comprehensive review. Clinical genetics. 2019;95(6):643-60.
234. Ma L, Chung WK. Quantitative analysis of copy number variants based on real-time LightCycler PCR. Current protocols in human genetics. 2014;80:Unit 7.21.
235. Salvi S, Conteduca V, Martignano F, Gurioli G, Calistri D, Casadio V. Serum and Plasma Copy Number Detection Using Real-time PCR. Journal of visualized experiments : JoVE. 2017(130).
236. Salvi S, Casadio V. Studying Copy Number Variations in Cell-Free DNA: The Example of AR in Prostate Cancer. Methods in molecular biology (Clifton, NJ). 2019;1909:95-103.
237. Mazaika E, Homsy J. Digital Droplet PCR: CNV Analysis and Other Applications. Current protocols in human genetics. 2014;82:7.24.1-13.

238. Stuppia L, Antonucci I, Palka G, Gatta V. Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases. International journal of molecular sciences. 2012;13(3):3245-76.
239. Schwarzenbach H. Copy Number Variation Analysis on Cell-Free Serum DNA. Methods in molecular biology (Clifton, NJ). 2019;1909:85-93.
240. Szuhai K, Vermeer M. Microarray Techniques to Analyze Copy-Number Alterations in Genomic DNA: Array Comparative Genomic Hybridization and Single-Nucleotide Polymorphism Array. The Journal of investigative dermatology. 2015;135(10):e37.
241. Magi A, Tattini L, Pippucci T, Torricelli F, Benelli M. Read count approach for DNA copy number variants detection. Bioinformatics (Oxford, England). 2012;28(4):470-8.
242. Magi A, Pippucci T, Sidore C. XCAVATOR: accurate detection and genotyping of copy number variants from second and third generation whole-genome sequencing experiments. BMC genomics. 2017;18(1):747.
243. Magi A, Bolognini D, Bartalucci N, Mingrino A, Semeraro R, Giovannini L, et al. Nano-GLADIATOR: real-time detection of copy number alterations from nanopore sequencing data. Bioinformatics (Oxford, England). 2019.
244. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. Genome research. 2014;24(12):2022-32.
245. Kader T, Goode DL, Wong SQ, Connaughton J, Rowley SM, Devereux L, et al. Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue. Genome medicine. 2016;8(1):121.
246. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nature communications. 2017;8(1):1324.
247. Abnizova I, Boekhorst Rt, Orlov YL. Computational Errors and Biases in Short Read Next Generation Sequencing. Journal of Proteomics & Bioinformatics. 2017;10(1).
248. Kono N, Arakawa K. Nanopore sequencing: Review of potential applications in functional genomics. Development, growth & differentiation. 2019;61(5):316-26.
249. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome biology. 2019;20(1):129.
250. Senol Cali D, Kim JS, Ghose S, Alkan C, Mutlu O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. Briefings in bioinformatics. 2019;20(4):1542-59.
251. Jeck WR, Lee J, Robinson H, Le LP, Iafrate AJ, Nardi V. A Nanopore Sequencing-Based Assay for Rapid Detection of Gene Fusions. The Journal of molecular diagnostics : JMD. 2019;21(1):58-69.
252. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nature methods. 2016;13(9):751-4.
253. Ni P, Huang N, Zhang Z, Wang DP, Liang F, Miao Y, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics (Oxford, England). 2019;35(22):4586-95.
254. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics (Oxford, England). 2018;34(18):3094-100.
255. Magi A, Semeraro R, Mingrino A, Giusti B, D'Aurizio R. Nanopore sequencing data analysis: state of the art, applications and challenges. Briefings in bioinformatics. 2018;19(6):1256-72.

256. Cheng SH, Jiang P, Sun K, Cheng YK, Chan KC, Leung TY, et al. Noninvasive prenatal testing by nanopore sequencing of maternal plasma DNA: feasibility assessment. Clinical chemistry. 2015;61(10):1305-6.
257. Dong Z, Xie W, Chen H, Xu J, Wang H, Li Y, et al. Copy-Number Variants Detection by Low-Pass Whole-Genome Sequencing. Current protocols in human genetics. 2017;94:8.17.1-8..6.
258. Pessoa LS, Heringer M, Ferrer VP. ctDNA as a cancer biomarker: A broad overview. Critical reviews in oncology/hematology. 2020;155:103109.
259. Webster TH, Couse M, Grande BM, Karlins E, Phung TN, Richmond PA, et al. Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. GigaScience. 2019;8(7).
260. Huw LY, O'Brien C, Pandita A, Mohan S, Spoerke JM, Lu S, et al. Acquired PIK3CA amplification causes resistance to selective phosphoinositide 3-kinase inhibitors in breast cancer. Oncogenesis. 2013;2(12):e83.
261. Liu P, Cheng H, Santiago S, Raeder M, Zhang F, Isabella A, et al. Oncogenic PIK3CA-driven mammary tumors frequently recur via PI3K pathway-dependent and PI3K pathway-independent mechanisms. Nature medicine. 2011;17(9):1116-20.